A Stochastic Context Free Model of Epistatic Interaction in the Human Genome

by

Jessica S. Schuster

B.A., Cornell University; Ithaca, NY, 2006

M.A., Brown University; Providence, RI, 2007

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2014

This dissertation by Jessica S. Schuster is accepted in its present form by The Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Charles Lawrence, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

William Thompson, Ph.D., Reader

Date_____ _____

James Padbury, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter Weber, Dean of the Graduate School

Vitae

Jessica S. Schuster

Applied Mathematics
Brown University
jess.schust@gmail.com
401-612-5020

Education

Ph.D *(Expected) May 2014*
Applied Mathematics, Brown University

- Advisor: Charles Lawrence
- Dissertation: "A Stochastic Context Free Model for the Inference of Epistatic Associations in the Human Genome"

M.S. *May 2008*
Applied Mathematics, Brown University

B.A. *May 2006*
Mathematics, Cornell University College of Arts and Sciences

- Advisor: John Guckenheimer

Research Interests

- Statistical Inference In High Dimensional Spaces and Statistical Modeling

- Computational Biology- Genome Wide SNP analysis, Linkage and Linkage Disequilibrium Detection, Population Stratification and Disease Association Studies
- Biology: RNAi and Somatic Mouse Cells

Academic Experience

Brown University: Teaching Assistant

- Statistical Inference
- Essential Statistics
- Methods of Applied Mathematics

Cornell University: Peer Tutor

- All areas Undergraduate Mathematics

Honors and Awards

- Brown University Dissertation Fellowship
- Brown University Fellowship
- Member of American Mathematical Association
- Member of Phi Beta Kappy Honor Society
- Member of Pi Mu Epsilon Mathematical Honor Society

Acknowledgements

There are several people I would like to thank, for without who I would not have been able to get to where I am at today. First and foremost, my sincerest gratitude goes to my advisor, Prof. Charles Lawrence for is support, guidance and patience throughout my graduate studies. He has taught, challenged and mentored me both academically and personally. His knowledge and visions have shaped my work at Brown and will continue to guide me throughout my career.

I want to thank the members of my lab group. Our weekly meetings have been both helpful for my own studies and thought provoking, introducing me the broad landscape that is Computational Biology. A special thank you goes to Prof. William Thompson for not only being a member of my dissertation committee, but for his infinite computer wisdom.

I also want to thank Prof. James Padbury for agreeing to be a member of my dissertation committee and for providing me with both the knowledge and data to be able to apply my algorithms to the study of preterm birth.

Finally I must thank my friends and family. You have always believed in me and stood by me through the good and the bad times. A most special thank you must be given to my husband, Andrew, for being my rock throughout this process and in all other aspects of my life. I look forward to what the future has in store.

Abstract of " A Stochastic Context Free Model of Epistatic Interaction in the Human Genome" by Jessica S. Schuster, Ph.D., Brown University, May 2014

Inheritable differences in DNA sequence are a major focus of the genetic community, as such differences are known to play a role in phenotypic variation, disease risk and environmental response. Single nucleotide polymorphisms account for a majority of human genetic variation and are advantageous to study due to their binary allelic nature, density in the genome and low rate of mutation. The study of the non-random association of the genotypes between two SNP loci, known as Linkage Disequilibrium, is of great interest as it is reflective of a wide range of population history, genetic subdivision, natural selection, epistasis and mutation and is integral in the study of disease association. Research on the block-like patterns of LD and the existence of LD over long genomic distance has produced a variety of conflicting results. Recent evidence has suggested the block-like patterns of LD are much for complex and that LD extends significantly further that previously believed. Current methods aimed at understanding and identifying patterns and long range LD and disease association fall short and suffer from computational limitations and complexity. This dissertation focuses on statistical inference in high dimensional space as it relates to epistatic interactions in the human genome. A stochastic context free grammar model is introduced as a novel method to infer associations in genome, yielding inference on jointly occurring pairs of pairs of associations both in close and distant proximity. PCA is used to access the existence of significant allele frequency differences between various ethnic groups. In the presence of such population stratification, the ancestry groups correlated to the top principal components are modeled by independent grammars, permitting the identification of population specific interactions. This statistical model allows for a complete search of the high dimension space of genomic data in a tractable amount of time, projecting the complex association landscape onto the constrained computable space of the context free model.

Contents

List of Tables

List of Figures

Chapter One

Motivation and Background

1.1 SNPs and Linkage Disequilibrium

Inheritable differences in DNA sequence is a major focus of the genetic community, as it is known that such differences play a role in phenotypic variation, disease risk and environmental response [1]. The source of much human genetic variation is in the form of Single Nucleotide Polymorphisms (SNPs), with additional, albeit less variation being the the form of insertions, deletions, rearrangements, and repeats. A SNP is a DNA variation where a single nucleotide in the genome differs between members of a species or between a pair of diploid chromosomes. SNPs have several properties which have made them advantageous to study. The binary allelic nature of SNPs make them suited for automated, high-throughput genotyping. In addition, they are relatively dense, occuring every 1,000-2,000 bases on average in humans and have a low rate of recurrent mutation[1]. Not long after sequencing the first human genome, a SNP map of the human genome was constructed [1]. This map consisted of approximately 1.42 million SNPs distributed across the genome, with an average density of about one SNP per 1.9 kilobases. The goal of producing such a map was to provide a public resource for understanding genetic variation and to ultimately identify the mutations contributing to disease.

While over a hundred genome wide association studies have been conducted on a wide variety of human diseases and many SNPs have been identified as related to disease risk, most common polymorphisms have been shown to have small effect size on disease risk. In certain cases, it has been hypothesized that a substantially large number polymorphisms would be required to account for the genetics of these complex disease. Goldstein suggests that the effect size might increase by considering interactions between genetic variants, but argues that it is still unlikely that these interactions would account for a majority of the variation [2]. However, the

role of epistatic interactions in phenotypic variation and disease risk should not be minimized. Epistatic interaction can occur both within and between molecules, with much of the mechanisms being poorly understood. For example, redundancy, both on the gene level and well as on the functional level, is know to produce epistatic interaction. Redundancy can be both full and partial, having a variety of effects on the phenotype as variations occur in both genes or pathways [3].

Covariation between polymorphisms may be studied as means to understand genetic variation and epistatic interactions. Although the SNP map was not released until 2001, the concept of non-random association of genotypes between two loci on a single chromosome, known as Linkage Disequilibrium, was introduced into the genetic community decades before [4]. Linkage disequilibrium is extremely important in human genetics due to the many factors that affect it and the many factored that are affected by it. Linkage disequilibrium throughout the genome may reflect population history and genetic subdivision, while region specific linkage disequilibrium is reflective of the history of natural selection, epistatis, mutation and gene conversion [5]. In addition, loci that are in close physical proximity are said to be genetically linked. Genetically linked loci are less likely to be separated to different chromatids during chromosomal crossover and thus appear to be inherited together. Chromosomal crossover is the exchange of genetic material between homologous chromosomes that results in recombinant chromosomes. In the absence of frequent recombination events, linkage disequlibrium results and these linked loci can appear as extended chromosomal regions with low diversity in allelic combinations. The allelic combinations are known as haplotypes and these regions are known as haplotype blocks.

Most commonly, linkage disequlibrium has been defined between alleles at two loci. There are several definitions of LD between the alleles at two loci that capture different aspects of the non-random association. At the base of all of these defini-

tions is the following: $D_{AB} = p_{AB} - p_A p_B$, which is the difference between the joint probability of the two loci haplotype AB and the product of the marginal probability of allele A at the first locus and the marginal probability of allele B at the second. The value D_{AB} is known as the disequilibrium coefficient and is defined for all combinations of alleles at the two loci. When D=0, the loci are in linkage equilibrium, which is similar to Hardy Weinberg Equilibrium (HWE) in that it reflects statistical independence between the two loci. Genotypes at a single locus are at HWE if the presence of an allele on one chromosome is independent of its presence on the homologue [5]. In addition, unlike HWE, disequlibrium decreases every generation at a rate dependent on recombination frequency [5]. Extensions to this coefficient have been made to make the value independent of allele frequency (D') [6] and to reduce the effect of the two loci having different allele frequencies by taking the ratio between D and the allele variance (r) [7]. The allele based linkage disequilibrium definition requires that an individual's gametes can be typed. However, for most studies, including the 1000 Genomes Project, the diploid genotypes are sequenced instead of the haplotypes. In such cases, the haplotypes must be statistically inferred from the genotypes.

There are several different approaches to this phasing problem, ranging from maximum likelihood to combinatorics. The downside of using statistically phased data to compute LD is that phasing ignores the uncertainty and possible errors inherent in the various inference algorithms [5]. To avoid such uncertainty from phasing inference, others have defined LD in terms of the diploid genotypes. Rogers and Huff [8] proposed an approximation method to calculate LD which collapses the eight parameter model of Weir [9] to a four parameter model. In this model, the correlation between the allelic values (r) is approximated by the correlation between the genotypic values and the allelic correlation coefficient can be inferred from the

approximation of r. Furthermore, Zhang et al., acknowledged the possible errors associated with haplotype phasing and instead used combinations of nearby diploid genotypes as they are observed in the data (diplotypes) to infer LD blocks and for association studies [10].

The size of haplotype blocks and the existence of LD over long distances has historically been the center of much debate. Prior to the production of the SNP map, some believed that LD was a direct predictor of physical distance [11, 12]. Computer simulations [13] and some empirical data analysis [14] concluded that LD around common SNPs extends only a few kilo-bases ($< 5kb$), while other empirical studies concluded that LD can extend over much greater distances ($> 100kb$). Such inconsistencies led to the large-scale study of Reich et al. [15] in which LD was studied in 19 randomly selected regions. Each of the 19 regions was centered at a core SNP with minor allele frequency of at least 35% in a multi-ethnic panel. Subregions of 2kb, centered at various distances extending to 160kb, were sequenced to identify common SNPs in which to measure disequilibrium. In comparing all 19 regions, the study found that LD blocks were generally larger than previously thought and confirmed that linkage disequilibrium generally decreases with the distance between markers. More importantly, they found great variation in LD amongst studied regions, including large variation in the length of the blocks over which LD extends as well as large variability in LD for SNPs in close proximity in regions where LD begins to drop off.

After several studies of various genomic regions finding non-overlapping haplotype blocks, separated by areas that displayed greater than normal rates of recombination, it was hypothesized that a majority of the human genome could be explained by this haplotype-block pattern, albeit with variation in the block length [16, 17]. This hypothesis gave much hope to the notion that each block would be able to be

represented by a single SNP and thus simplifying the space to test for disease association [18]. Several methods for inferring haplotype blocks over a given genomic region can be found in the literature. There are two major types of blocking methods, one in which blocks are based on the decay of LD between blocks[16, 17, 19] and the other is based on haplotype diversity within blocks [20, 21]. Daly [16] used a Hidden Markov Model in which transitions between states is related to the decay of LD measured by D', whereas Wang [19] used the Four Gamete Test (FGT) of Hudson & Kaplan (1985) to detect boundaries based on decaying LD. The FGT detects recombination events by locating pairs of SNPs that cannot have arisen without either recombination or a repeat mutation. Gabriel [17] defined blocks based on the proportion of marker pairs showing evidence for historical recombination based on an arbitrarily defined upper bound of the 95% confidence interval of the D' statistic. These three algorithms fall into the first category of blocking methods as they are all founded in the detection of decaying LD between blocks. The algorithms by Patil [20]and Zhang [21] fall into the second category as they are based around finding blocks the with low common haplotype diversity and the belief that most common haplotypes within blocks can be identified using a small number of SNPs. The smallest group of SNPs needed for identification of the common haplotypes are known as tagging SNPs. Patil implements a greedy algorithm and Zhang implements a dynamic programming algorithm to find the block partition which has the globally minimum number of SNPs.

Anderson & Novembre proposed an algorithm which combines both decay of LD between blocks and limited haplotype diversity within blocks to define block partitions. This algorithm uses the principle of Minimum Description Length (MDL) to select block designations. The MDL principle is an application of information theory to statistical model selection, in which the set of block boundaries is chosen

such that they correspond to the shortest description length for the data[22]. A comparison of the results from MDL method with the results from several of the blocking algorithms above, concluded that while the results using the methods of Jefferys [23]and Cullen [24] suggest recombination hotspots are associated with the blocks of LD and the results of Gabriel [17] and Kauppi [25] further suggest that these recombination hotspots are in fact at the block boundaries, the results using the methods of Wang [19] indicated block boundaries exist at both hot-spots and non-hotspot locations. In addition, they found the Wang's FGT method produces a significantly different set of blocks when compared with their MDL method and the HMM method of Daly. Via simulations, they demonstrated that the MDL method could in fact detect blocks in the absence of hotspots [22]. Given al of these contradictory results, it is evident that partitioning the genome into blocks is not only a non-trivial task, but is insufficient in understanding the complete landscape of LD.

In addition, one aspect of LD that is not fully addressed by current haplotype blocking algorithms is long range LD that extends beyond the range of recombination and thus can exist between non-contiguous regions of the genome. Some recent attempts have been made to identify and understand long range LD. The works of Koch et al. attempts to both identify and characterize genome-wide patterns of long range LD by computing all pairwise LD via the Fisher exact test. In addition to these pairwise measurements, they apply a heuristic procedure to define patches, i.e. regions of SNPs that are statistically associated to other distant regions of SNPs. Two pairs of SNPs are considered to be in a patch if each pair of distant SNPs has a high LD measurement and if the distance between the left loci of the two pairs, as well as the distance between the right loci of the two pairs are less than some specified value. Additional pairs are added to the patch if they meet the criteria when compared to at least one other pair already in the patch. They further look

for non-random patterns by testing for statistical significance of the LD max and the number of patches by chromosome with Bonferoni correction for tests of multiple chromosomes [26]. In the sections to follow, a new method will be presented to infer SNPs and regions of SNPs in long range LD jointly via a fully probabilistic context free grammar. As results from the probabilistic context free grammar are a mathematical consequence of the model and the data, there is no need to use heuristics for the inference pairs of pairs of associated SNPs.

Furthermore, in addition to the variability in the extent to which linkage disequilibrium exists across regions of the genome and the inconsistencies in defining haplotype blocks, there exists variability in linkage disequilibrium between ancestry groups. In the aforementioned study of Reich et al., individuals of northern-European ancestry and Nigerian ancestry were genotyped. The half-life of LD extended to 60kb on average for common SNPs in the northern-European sample and exhibited very similar pattern to the sample from Utah, while LD was present at significantly shorter distances in the Nigerian population. This study found very little Nigerian specific LD and attributed the differences in the distance that LD extends to a European bottleneck [15]. Some other ancestry specific studies also detected longer than expected LD blocks explained by bottleneck, as well as a subdivision of the African population followed by recent admixture [27, 28]. This variability in LD can be attributed to presence of allele frequency differences populations of differential ancestry. This allelic difference is know as population stratification. Identification of population stratification is extremely important since it can result in the presence of spurious associations between SNPs.

Several methods have been developed to deal with population stratification, including genomic control and structured association. Genomic control corrects for stratification by uniformly adjusting the association statistics for all pairs of mark-

ers by an inflation factor [29]. This approach suffers from a loss of power because the extent of allele frequency differentiation is not constant over the genome. Structured association divides sampled individuals into mutually exclusive clusters and preforms association inference on each cluster [30]. This method is sensitive to both number of clusters included in the analysis and the division of samples into clusters. In 2006, Price et al. [31] proposed a method that would address the aforementioned limitations while identifying and correcting for stratification. At the backbone of this method named EIGENSTRAT, is Principle Component Analysis. PCA is used to infer axes of genetic variation. After these axes are defined, the genotype data can adjusted along each of the significant axes, which is comparable to computing the residuals of a linear regression. Association analysis can be performed on this adjusted data. In the following chapter, as LD is studied in geographically diverse sampled populations, PCA is use to infer the appropriate subpopulations for independent analysis.

As new algorithms were developed and more and more genomic regions were tested, it became increasingly evident that more work needed to be done to fully understand linkage disequilibrium and the haplotype block pattern of the human genome. In response, the International HapMap Project was formed. In the first phase of the project, more than 1 million SNPs were genotyped and LD was measured in 269 individuals from four diverse populations (European, Chinese, Japanese, and Yoruban) [32], In the second phase, 3.1 million SNPs were genotyped from the same sample population [33]. Increased efficiency and cost effectiveness prompted larger scale genotyping projects. In 2008, the 1000 Genomes Project got underway in an effort to sequence over 1000 individuals from a wide array of ethnic groups in order to provide a comprehensive characterization of variation in the human genome [34].

With the new sequencing technologies and an influx of polymorphic genotype

data, Zhang & Liu [10] introduced a bayesian epistatic association mapping algorithm, BEAM, for inferring marker interactions and disease associations. Their algorithm used a Markov Chain Monte Carlo method with Metropolis Hastings sampling to evaluate each marker conditional on the current classification of all other markers and outputs an approximation of the posterior probability of each markers classification, i.e. independent or epistatic. The posterior probability of the the classification vector given the data is proportional to the product of several Dirchlet-Multinomial probabilities, with each state/classification having a different specification of the Dirchlet-Multinomial. A MCMC is used to approximate this posterior distribution conditional on the data because the normalizing constant is too complex to calculate [10]. While this method was shown to perform as well or better than previous association algorithm, the MCMC approach is only an approximate representation of the posterior space and there are the problems of the increased time and of assessing convergence as all states will not be visited [35].

In the remaining chapters, a new approach will be introduced, using a Stochastic Context Free Grammar as a model to infer multiple associated pairs in both close and distant proximity. The Stochastic Context Free model is a fully defined probabilistic model in which the partition function of the posterior distribution can be calculated in $O(n^3)$ time and space. In addition, unlike many existing algorithms, it does so without restraint on window size and multiple comparison error/correction. Similar to Zhang et al. [10], in order to avoid to uncertainty in regards to haplotype phasing, linkage disequlibrium will be investigated at the unphased diplotype genotype level via the sampled correlations. The model can be easily adjusted to incorporate prephased haplotype data sequences.

To follow, Chomsky's Natural Grammars are introduced and the Stochastic Context Free Grammar model of linkage disequilibrium for SNP genotype data is pre-

sented, along with the probabilistic algorithms used for learning model parameters and making inference on the LD landscape. The algorithm is then applied to both simulated data as well as two real data sets taking into account population stratification, with results to follow. Future directions in which the model may be extended would allow for more complex structures of LD and application of the model to disease association studies are discussed.

1.2 The Stochastic Context Free Grammar

1.2.1 Chomsky Grammars: Regular, Context Free & Stochastic

A grammar, G, is a set of production rules for strings of a formal language, $L(G)$. The rules describe how to form strings from the language alphabet that conform to the constraints of the language syntax. Chomsky proposed that all generative grammars consist of four parts: (1) finite set of non terminal states (N) which do not appear in strings, (2) set of terminal states (Σ) that are disjoint from N, (3) set of production rules (P) and (4) the start symbol ($S \in N$). The production rules for the general grammar, G, are of the form: $(\Sigma \cup N)^\star N(\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*$ where * is the Klene star operator. The Klene star is a unary operator such that the Klene star on set V (V^*) is the collection of all possible finite-length strings generated from the symbols in set V. Grammars can be viewed as language generators in that a string of the language can be derived from a grammar by a sequence of rule applications beginning at the non-terminal start state S ending in a string of terminals: $S \Rightarrow s_1 \Rightarrow s_2 \Rightarrow s_3 \Rightarrow \ldots \Rightarrow x$, where $s_i \in (\Sigma \cup N)^\star$and $x \in \Sigma^\star$.

We denote such a derivation of x as $S \stackrel{\star}{\Rightarrow} x$. The existence of a derivation of a string of terminals x also proves that the string belongs to the grammar language, thus grammars can also be viewed as language recognizers and the language of a grammar is the set: $L(G) = \{x \in \Sigma^* : S \stackrel{\star}{\Rightarrow} x\}$ [36].

Chomsky defined a hierarchy of grammars that impose various restrictions on the production rules and the types of languages that can be produced. One class is known as Regular Grammars. A regular grammar is a grammar such that the left hand side of all of the production rules are a single non-terminal state and that the right hand side of all production rules can be of three types: the empty string, a single terminal, and a single terminal with a single non terminal where the terminal is on the left or right of the non-terminal for all such productions [36].

The second class of grammars is known as the Context Free Grammars. Like a regular grammar, a context free grammar restricts the productions rules such that the left hand side of all production rules is a single non-terminal state; however, there are no restrictions on the right hand side of the grammar: $X \rightarrow \lambda$ where $X \in N$ and $\lambda \in (\Sigma \cup N)^*$. A regular grammar is by definition a context free grammar; however, not all context free grammars are regular grammars[36]. An example of a context free grammar, $G1$, that is not a regular grammar, has the following set of production rules:

$$\begin{aligned} S &\rightarrow A \\ S &\rightarrow bSbb \\ A &\rightarrow aA \\ A &\rightarrow \epsilon \end{aligned}$$

where $N = \{S, A\}$, $\Sigma = \{a, b\}$ and the start state is S. The language for this

grammar is: $L(G1) = \{b^n a^m b^{2n} : n \geq 0, m \geq 0\}$.

As demonstrated above, a context free grammar can produce a pair of terminals from a single non-terminal; however, a regular grammar requires two different non-terminals to produce each terminal of a pair independently. In addition, regular grammars generate strings from left to right or vice versa, while context free grammars generate strings from the outside-in. This outside-in generation restricts paired terminals to be nested. Consider the following grammar, $G2$:

$$
\begin{aligned}
S &\rightarrow SS \\
S &\rightarrow () \\
S &\rightarrow (S) \\
S &\rightarrow [] \\
S &\rightarrow [S]
\end{aligned}
$$

where ( ) and [ ] are terminal states and S is the only non-terminal state. This grammar does not generate all strings with two different types of parenthesis. For example, it generates the string ([[[()]]]) where all parenthesis are paired and nested, but it does not generate ((([])))) since there is a conflicting pair of [] and (). The context free grammar differs from the third class of grammars, known as Context Sensitive grammars, in that the the single non-terminal on the left hand side of the productions rules prohibits derivations from being dependent on it neighbors [36].

Regular and context free grammars dictates whether a string of terminals x is a member of the grammar's formal language or not. These two classes of grammars can be extended such that they are probabilistic and thus each string of terminals can also have a associated probability of being generated by the grammar according to

the parameter space θ. This probabilistic model is referred to as a stochastic grammar. In a stochastic grammar, for a given non-terminal state X, the probabilities of all possible productions rules from state X sums to one ($\sum_{\lambda:X\to\lambda} Prob(X \to \lambda) = 1$) and the stochastic grammar defines a probabilistic distribution over all strings of the language ($\sum_{x\in L(G)} Prob(x|\Theta) = 1$)[36]. A Hidden Markov Model can be viewed as the stochastic regular grammar and the Stochastic Context Free Grammar (SCFG) is the probabilistic version of the context free grammar. Below are two applications of SCFG models, which provide the motivation for its use studying epistatic associations.

1.2.2 SCFG and Natural Language

Historically, Stochastic Context Free Grammars have been used to model languages spoken by the human population [37]. For Natural Languages, words maybe have multiple meanings and the same sentence can have multiple interpretations, and the probabilistic grammar models this complexity of by allowing for more than one set of production rule applications to produce a single sequence of words along with their relative weights. The following grammar is an example from Manning and Schtze [38]:

$$S \rightarrow NP\,VP \quad 1.0$$

$$PP \rightarrow P\,NP \quad 1.0$$

$$VP \rightarrow V\,NP \quad 0.7$$

$$VP \rightarrow VP\,PP \quad 0.3$$

$$P \rightarrow with \quad 1.0$$

$$V \rightarrow saw \quad 1.0$$

$$NP \rightarrow NP\,PP \quad 0.4$$

$$NP \rightarrow astronomers \quad 0.1$$

$$NP \rightarrow eyes \quad 0.18$$

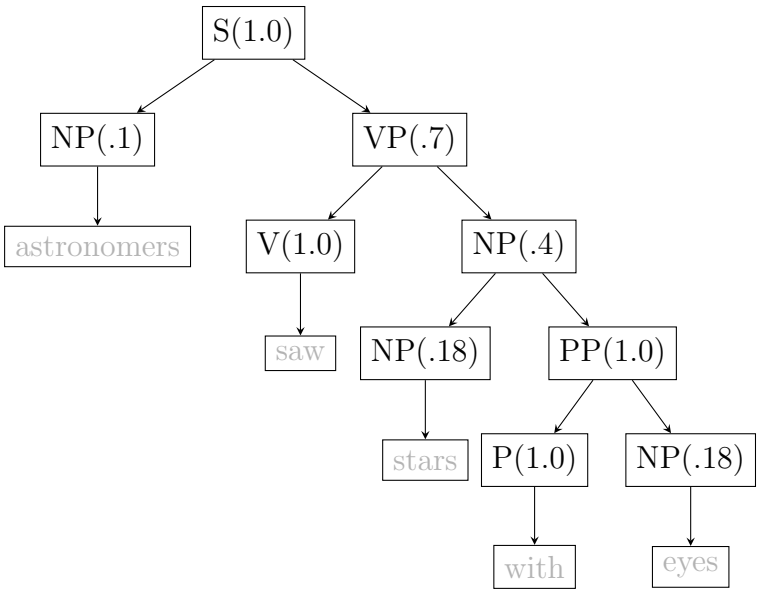
$$NP \rightarrow saw \quad 0.04$$

$$NP \rightarrow stars \quad 0.18$$

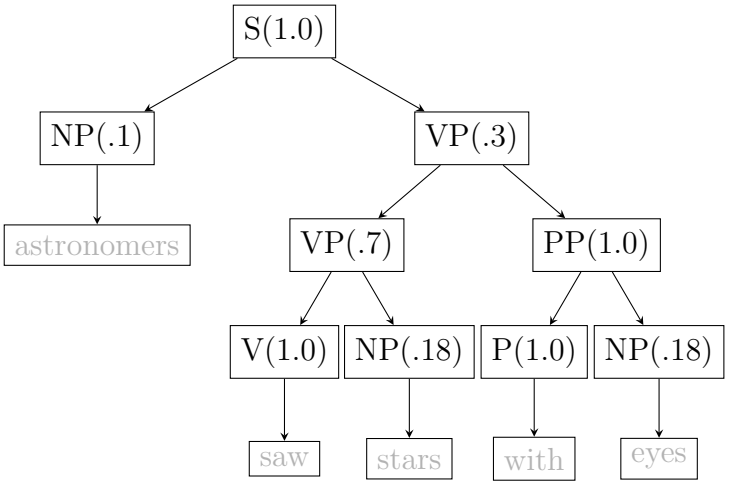
$$NP \rightarrow telescopes \quad 0.1$$

The string of words, "astronomers saw stars with eyes", has two possible interpretations: (1) *astronomers used eyes to see the stars* or (2) *the stars that the astronomers saw had eyes.* The difference between these two interpretations is whether the prepositional phrase PP "with eyes" is grouped with "saw stars" or just "stars". These interpretations can each be derived from a different parse tree as shown in Figure 1.1.

Since more than one possible parse tree exists for a single string of words, this grammar is known as *ambiguous* [39]. In order to model the English lan-



(a) T1: astronomers saw stars with eyes



(b) T2: astronomers saw stars with eyes

Figure 1.1: *"Astronomers saw stars with yes"*. According to the rules of the grammar, there are two possible parse trees that exist for this string of words, each with a different meaning. (1) *astronomers used eyes to see the stars* or (2) *the stars that the astronomers saw had eyes.*

guage it is necessary for the grammar to be *ambiguous* to capture its complexity and the stochastic component allows for weights to be assigned to the various parses/interpretations as demonstrated below:

The probability of the sentence given parse $T1$ and the model parameters θ can be calculated as follows:

$$Prob(w, T1|\theta) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = .0009072$$

and the probability of the second derivation $T2$:

$$Prob(w, T2|\theta) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = .0006804$$

In addition, the SCFG generates all the valid sentences of a language and provides a means to access the probability of a sentence being generated by the model/ being part of the language as seen below:

The probability of sentence w="astronomers saw stars with eyes" can be calculated as follows:

$$Prob(w|\theta) = Prob(w, T1|\theta) + Prob(w, T2|\theta)$$

1.2.3 SCFG and RNA Secondary Structure Prediction

A generalization of the SCFG has also been used to model RNA secondary structure [40, 41]. RNA molecules are involved in many biological processes and their structure is a determining factor in their function (Moore, 2003). For example, the structure of messenger RNA affects post transcriptional regulation including alternative splicing

control [42] and the regulation of RNA stability [43]. However, it is extremely difficult and time consuming to determine the structure of RNA using crystallization or other experimental procedures. Therefore, researchers have been focused on understanding the secondary structure of an RNA, which is is believed to be a driving element behind the three dimensional folded RNA structure. The secondary structure is formed from complementary base pairing of the nucleotides of the molecule. Two secondary structures are considered identical if they have have same set of base pairs. The full ensemble of secondary structures for a given RNA sequence is of great interest to biologists [44, 41] as certain RNA may exist as a population of different structures [45] and multiple structures of an RNA are known be involved in a variety of functions [46, 23, 47, 48].

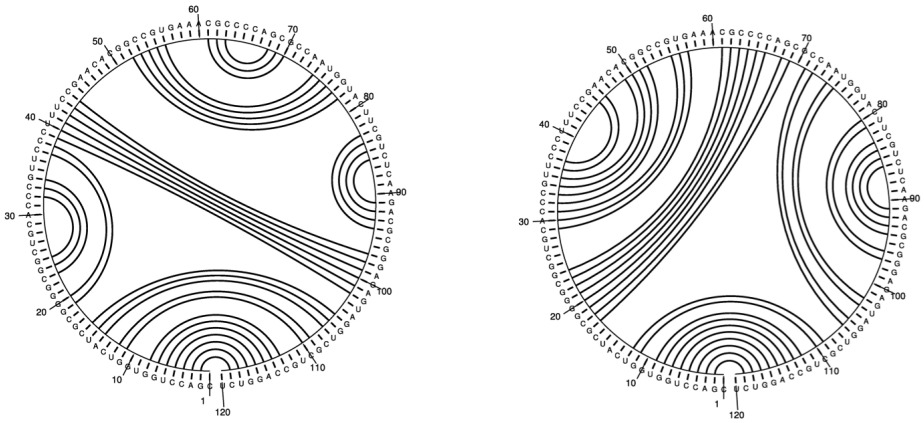
The use of the SCFG in RNA secondary structure prediction was motivated by the use of HMMs in protein modeling [49]. Unlike the HMM, the SCFG allows for long range interactions and thus the have been used to model the nested complementary nucleotide pairs in RNA [50, 51]. In addition, just as the probabilistic grammar model of natural language above produces two different interpretations of the sentence "astronomers see stars with ears", ambiguous stochastic grammars have been implemented to model all possible un-pseudoknotted secondary structures for a given RNA sequence of nucleotides and their associated probabilities [52]. It is important to note that while the SCFG for RNA is ambiguous in this sense, it must also be *structurally unambiguous*, that is there must exist a one to one correspondence between the parse trees and the secondary structures. A symmetric matrix of zeros and ones, indicating the presence of a base pair can be used to represent the secondary structure of RNA and structural unambiguity ensures that the probability of a parse tree will be equal to the probability of the secondary structure.

In its simplest form, a SCFG model can produce an ensemble of totally comple-

mentary RNA sequences. Several programs, including Pfold [53] and CONTRAfold [54] have used more complex SCFGs to model single stranded RNA regions, stacking of complementary pairs and other complex biological features. Other packages such as Sfold [16], RNAstructure [55], and the Vienna Package [56] have also been used model secondary structure using their thermodynamic properties. Two common features amongst all of these models are the following constraints on possible nucleotide pairs: (1) each base can pair with at most one other base and (2) pairs cannot conflict, i.e. if SNP i and j are paired, there cannot exist a pair (k, l) such that $i < k < j < l$ or $k < i < l < j$ for all pairs (i, j).

In their 2003 paper, Ding and Lawrence introduced an algorithm to back-sample exactly from the ensemble of constrained secondary structures and the authors demonstrated that the sampled structures are statistically representative. While it is not computationally feasible to enumerate all the possible structures for an RNA sequence, the ensemble can be split into mutually exclusive classes and all classes except those whose contribution to the posterior probability mass are very small will almost certainly be represented in the sample. Ding and Lawrence demonstrate that via statistical sampling and clustering, the probability of any structural motif can be estimated [40]. Ding and Lawrence used their algorithm to predict the secondary structure of 5S RNA. For each class of structures the centroid estimator, i.e. the structure which minimizes the expected hamming distance amongst all structures in the class, was determined [57]. Figure 1.2b contains the centroid estimators of the two main classes sampled from the ensemble of structures for the 5S RNA.

These two classes of structures formed as a result of constraint violations. For example, in Figure 1.2a , base 61 is paired to base 70, whereas in Figure 1.2b, base 24 is paired to base 61. These pairings produce two classes of structures due to the constraint that each base can pair to at most one other base. In addition, in Figure



(a) Centroid Cluster 1 **(b)** Centroid Cluster 2

Figure 1.2: *5S RNA*. The centroids for the two clusters with the highest probability mass. Two classes formed due to pairs which violate of the constraints that each and that contain that if SNP i and j are paired, there cannot exist a pair (k, l) such that $i < k < j < l$ or $k < i < l < j$ for all pairs (i, j).

1.2a base 40 is paired to base 101, whereas in 1.2b, base 24 is paired to base 61. These two pairs cannot appear in a single structure due to the constraint if SNP i and j are paired, there cannot exist a pair (k, l) such that $i < k < j < l$ or $k < i < l < j$ for all pairs (i, j). Although the grammar cannot predict a structure where base 61 pairs with 70 and 24, or both pairs (24-61) and (40-101), sampling identifies these possible pairings. It is possible to identify pairs of pairs that violate the constraints listed above and sampling provides a greater understanding about all structures in the ensemble. Multiple such classes of structures can thus be seen as a projection of the high dimensional space that lacks these constraints (space of pseudoknotted structures) into a space that does have the constraints.

1.2.4 SCFG: The most probable parse and the partition function

As demonstrated by the two applications above, the Stochastic Context Free Grammar is a natural model for structured sequence data, specifically sequences with matched elements. When presented with a set of sequential data, there are several questions of interest: (1) what is the probability that grammar, G, generates the string x? (2) which parse of the string x is most probable with respect to the grammar? (3) given a set of data sequences, how can the model parameters, Θ, be estimated? The power of using Stochastic Context Free Grammars to model sequential data is that these questions can answered using recursive dynamic-programming-like algorithms.

What is the probability that grammar G generates sequence x? The probability of interest can be computed with a generalization of the *Forward Algorithm* for HMM's, known as the *Inside Algorithm* [58]. The *Inside Algorithm* calculates the probability of generating any substring given the substring is rooted in any non-terminal state $X \in N$ and grammar, $G = \{N, \Sigma, P, S, \theta\}$ [58]. Since the grammar derives strings of terminals from the outside-in, the algorithm starts with substrings of length 1 and works outwards, capitalizing on the previously calculated probabilities, until the probability of the whole sequence is calculated. The joint probability of the data given the model and its parameters is known as the partition function.

With respect to the grammar G, which derivation/parse of the data sequence x is most probable? The *Cocke–Younger–Kasami (CYK) Alignment Algorithm* can be used to find the most probable derivation [59]. This algorithm is a variation of the *Inside Algorithm*, replacing the sum over all possible derivations by the maxi-

mization function. This is an outwardly growing recursive algorithm that calculates the probability of the maximal derivation for any substring given that the substring is rooted any non-terminal state $X \in N$. While calculating the probability of the most likely derivation, a traceback algorithm is implemented to recover the parse of interest.

Given a set of data sequences, how can the parameters ,Θ, of the grammar, G, be estimated? The generalized of the *Baum-Welch Algorithm* for HMMs, known as the *Inside-Outside Algorithms* can be used estimate the maximum likelihood parameters θ of the SCFG via an iterative *Expectation-Maximization Algorithm* [36]. In addition, samples can be drawn directly from the joint distribution of the ensemble of all possible derivations of the sequence data, where the joint distribution is calculated via the *Inside Algorithm*. From a representative sample of derivations, the marginal probabilities of matched pairs can be computed.

1.2.5 SCFG and Genome-Wide Linkage Disequilibrium

Historically, most efforts to identify Linkage Disequilibrium focus on inference based on individual pairs of SNPs or local regions of genome sequences. The goal of this research is to gain additional power in assessing linkage disequilibrium genome wide through the joint distribution of the genotype data of pairs of pairs of SNPs, using a Stochastic Context Free Grammar. To model the association between SNPs, the grammar will contain a non-terminal state that emits a pair of associated SNPs, as well as a non-terminal state that emits an unassociated/unpaired SNP. As with the case of RNA secondary structure, this model imposes the following constraints on the derivations of the sequence of SNPs: (1) each SNP can be associated to at most one other SNP, (2) if the SNP at position i is paired to the SNP at position

j, there can not exist a pair of associated , snps k and l such that $i \leq k \leq j \leq l$ or $k \leq i \leq l \leq j$. Neither of these constraints are biologically motivated, but rather they are limitations of the context free grammar. As with RNA secondary structure, sampling from the ensemble of parse trees will allow identification of pairs of SNPs that violate these constraints and thus will allow us to make inference on a more complex landscape.

Chapter Two

The Data, The Model and The Algorithms

2.1 The Data

For a polymorphic data set, let M represent the number of individual genotypes sampled and let L represent the number of SNP loci sequenced for each individual. For individual m, let the genotype of SNP j be represented as $s_{m,j} = \{0, 1, 2\}$. The integer value of the genotype represents the number of alternative alleles at the locus. $S_{<1,L>}$ is the $M \times L$ matrix of genotypes $s_{m,j}$ and $S_{<i,j>}$ is the matrix of genotype data for a substring of SNPs from the i^{th} position to the j^{th} position.

2.2 The Model: Stochastic Context Free Grammar G

The Terminal State Space

Inference on the joint probability of the states of all SNPs will be drawn by collapsing individual genotype data into genotype counts for single and paired loci. Thus, the terminal state space, $\mathbf{\Sigma}$, consists of the counts of the genotypes of the SNPs. The terminal states take two forms: (1) the summary statistic for all single SNPs i, S_i and (2) the summary statistic for all pairs of SNPs at positions i and j, S_{ij}.

The M dimensional vector of genotypes for SNP i, $\vec{s_{\bullet i}}$, can be collapsed to the 3×1 vector of genotype counts, S_i. Let $n_{k^i} = \sum_{m=1:M} I_k(s_{mi})$ where $I_k(s_{mi}) = 1$ if $s_{mi} = k$ and $I_k(s_{mi}) = 0$ otherwise. Thus n_{k^i} is the number of times genotype k appears at SNP locus i and the count vector is $S_i = (n_{0^i}, n_{1^i}, n_{2^i})$.

In addition, the $M \times 2$ dimensional matrix of genotypes for SNP pair i and j can be collapsed to the the 3×3 matrix of genotype counts S_{ij}. Let $n_{k^i l^j} = \sum_{m=1:M} I_k(s_{mi}) \times I_l(s_{mj})$ where $I_k(s_{mi}) = 1$ if $s_{mi} = k$ and $I_k(s_{mi}) = 0$ otherwise and $I_l(s_{mj}) = 1$ if $s_{mj} = l$ and $I_l(s_{mj}) = 0$ otherwise. Thus $n_{k^i l^j}$ is the number of individual with genotype k at position i and genotype l at position j and the matrix of genotype counts is:

$$S_{ij} = \begin{pmatrix} n_{0^i 0^j} & n_{0^i 1^j} & n_{0^i 2^j} \\ n_{1^i 0^j} & n_{1^i 1^j} & n_{1^i 2^j} \\ n_{2^i 0^j} & n_{2^i 1^j} & n_{2^i 2^j} \end{pmatrix}$$

.

For a small example, let $M = 5$ and $L = 7$, i.e. 5 individuals with 7 genotyped SNPs. Table 2.1 contains the genotype matrix $S_{<1,7>}$. There are two individuals (row 1 and 3),with genotype 0 for SNP 1 (the first column of the table), so $n_{0^1} = 2$. Likewise, for SNP 1, there are two individuals (row 2 and 4) with genotype 1, so $n_{1^1} = 2$ and there is one individuals with genotype 2 (row 5), so $n_{2^1} = 1$. Thus, the count vector for SNP locus 1 is $S_1 = \{2, 2, 1\}$. Similarly, the count vectors for the remaining SNPs in Table 2.1 are:

$$\begin{aligned} S_2 &= \{3, 2, 0\} \\ S_3 &= \{2, 1, 2\} \\ S_4 &= \{2, 2, 1\} \\ S_5 &= \{2, 2, 1\} \\ S_6 &= \{2, 2, 1\} \\ S_7 &= \{2, 2, 1\} \end{aligned}$$

ind/snp	1	2	3	4	5	6	7
1	0	0	2	1	1	0	1
2	1	0	0	0	1	1	1
3	0	1	0	2	0	0	0
4	1	1	2	1	0	2	0
5	2	0	1	0	2	1	2

Table 2.1: *An example of Genotype Data* Genotype data for 5 individuals and 7 SNPs. Genotypes take values of 0,1, 2 corresponding to the number of minor alleles at the loci

In addition, there is one individual (row 1) with genotype 0 for the SNP 1 AND genotype 0 for the SNP 2, so $n_{0^1 0^2} = 1$. Likewise, $n_{0^1 1^2} = 1$, $n_{0^1 2^2} = 0$, $n_{1^1 0^2} = 1$, $n_{1^1 1^2} = 1$, $n_{1^1 2^2} = 0$, $n_{2^1 0^2} = 1$, $n_{2^1 1^2} = 0$, and $n_{2^1 2^2} = 0$. Thus, the count matrix for the pair of SNPs 1 and 2 is $S_{12} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$. This statistic can be computed for all possible pairs of SNPs.

The Non-Terminal State Space

The second component of the Stochastic Context Free Grammar, $\boldsymbol{G}$, is the non-terminal state space, $\boldsymbol{A}$. The non-terminal states define the underlying structure of the language. The-non terminal states and their definitions are as follows:

- S: generates any string of SNPs and their genotype counts
- N: generates any non-empty string of SNPs and their genotype counts
- E: generates the empty string
- L: generates strings with the left most SNP associated to exactly one other SNP

- P: generates strings with the left most SNP associated to the right most SNP
- U: generates strings with all the SNPs unpaired/unassociated
- UL: generates bifurcated strings with the substring of SNPs up to and including the bifurcation point generated by state U and the substring of SNPs to the right of the bifurcation point generated by state L
- PS: generates bifurcated strings with the substring of SNPs up to and including the bifurcation point generated by state P and the substring of SNPs to the right of the bifurcation point generated by state S
- PN: generates bifurcated strings with the substring of SNPs up to and including the bifurcation point generated by state P and the substring of SNPs to the right of the bifurcation point generated by state N.

The Production Rules

The third part of the grammar is a set of production rules $\boldsymbol{R}$ which define legal transitions between states and thus dictate the possible parse trees. For the stochastic context free grammar, all production rules are of the following form: $V \rightarrow \lambda$ where $V \in \boldsymbol{A}$ and $\lambda \in (\Sigma \cup \boldsymbol{A})^*$.

The non-terminal state S, which can generate all strings, can produce either the state generating non empty strings, N or the state generating the empty string, E ($S \rightarrow N|E$) The "|" symbol indicating that the non-terminal on the left hand side of the production rule can produce either state on the right hand side of the arrow.

The non-terminal state N, generating all non-empty strings, can produce three non-terminal states ($N \rightarrow L|UL|U$) which generate three categories of strings: (1)

strings with the leftmost position paired (2) bifurcating string with the first k elements from a substring generated by state U and the remaining elements a substring generated by state L and (3) string with the all unpaired elements, respectively.

The non-terminal state L, which generates all strings with the left position paired, produces the non-terminal state PS ($L \rightarrow PS$). This bifurcating state PS produces a left substring generated by state P and a right substring generated by S ($PS \rightarrow P/S$). Similarly, the bifurcating state PN produces a left substring generated by state P and a right non-empty substring generated by N ($PN \rightarrow P/N$) and UL produces a left substring generated by U and a right substring generated by L ($UL \rightarrow U/L$). The "/" symbol indicating that the bifurcating non-terminal on the left hand side of the production rule produces substrings rooted in both non-terminals on the right hand side.

All of the above productions have been from a non-terminal state to a non-terminal state or a pair of non-terminal states. Non-terminal states P and U are emitting states, which means they produce a string of terminal and non-terminal states. State U emits a terminal state (unpaired summary statistic) and produces the non-terminal state U, that generates a substring one SNP shorter then the state which produced it or produces the non-terminal state E, which generates the empty string ϵ. ($U \rightarrow U|E$).

Similarly, state P emits a pair of terminal states (paired summary statistic) and produces one of the following non terminal states ($P \rightarrow P|U|PN|UL|E$) which generate strings that are two elements shorter than the states that produced it.

All of the production rules, $\boldsymbol{R}$, for the grammar $\boldsymbol{G}$ are summarized as follows:

$$
\begin{aligned}
S &\rightarrow N \mid E \\
N &\rightarrow L \mid UL \mid U \\
L &\rightarrow PS \\
PN &\rightarrow P/N \\
PS &\rightarrow P/S \\
UL &\rightarrow U/L \\
U &\rightarrow xU \mid xE \\
P &\rightarrow yPy \mid yUy \mid yPNy \mid yULy \mid yEy \\
E &\rightarrow \epsilon
\end{aligned}
$$

where x and y are the terminal-state states for the unpaired positions and the paired positions respectively.

The Probability Distribution for $\boldsymbol{R}$

For all non-terminals in $\boldsymbol{A}$, there is a normalized probability distribution on the production rules $V \rightarrow \lambda$. The probability distribution can be broken into to parts: the *transition probabilities* and the *emission probabilities* The transition probability is the probability of the production from one non-terminal state to another. For notational purposes, the transition probability for production rule $V \rightarrow Y$ is $t(V, Y)$. The transition probabilities follow a multinomial distribution such that $\sum_Y t(V, Y) = 1$, where V and Y are non-terminal states.

Non-terminal states P and U are known as emitting states, i.e. they produce

both a non-terminal/internal node and an terminal/external leaf of the parse tree simultaneously. For these two states, the probability of the production rule $V \rightarrow \lambda$ is the product of the corresponding transition probability and the emission probability, $emit(x|V)$. The emission distribution is defined in Section 2.3.

An Example: The Grammar In Action

This grammar $\boldsymbol{G}$ is *structurally unambiguous*, meaning there is one to one correspondence between the the parse trees and matrix of associated pairs of SNPs (aka the LD structure) for a given sequence. This one to one correspondence is necessary for the inference of the most probable structure for a given genotype sequence, and to calculate the joint probability of the genotype sequence and the location of the pairs, given the model parameters. This probability is known as partition function and will be computed in Section 2.4

Figure 2.1 contains the unique parse tree for the toy example from the previous section that produces the structure where only loci 4 and 6 are associated. The parse tree, like all other parses, starts at state S. State S transitions to state N. State N then transitions to bifurcating state UL. UL transitions to the two non-terminal states U and L, with the former generating the substring $S_{<1,3>}$ and the latter generating $S_{<4,7>}$. Non-terminal state U emits terminal S_1 and produces non-terminal state U which generates $S_{<2,3>}$. State U, which is the root of $S_{<2,3>}$, emits terminal S_2 and produces non-terminal state U. This non-terminal, which generates $S_{<3,3>}$, again emits terminal S_3 and transitions to the empty string state. Simultaneously, non-terminal state L which is the root of $S_{<4,7>}$ transitions to bifurcating state PS, that subsequently transitions to the two non-terminal states P and S. P is the root

of $S_{<4,6>}$ and S is the root of $S_{<7,7>}$. State P emits the paired terminal state $S_{4,6}$ and produces state U, which emits terminal state S_5 and transitions to the empty string state. Finally, state S, which is the root of $S_{<7,7>}$, transitions to N, which transitions to U and state U emits the final unpaired terminal state S_7. This the the unique parse tree that produces a structure in which SNPs 4 and 6 are paired.

Example of the the SCFG in action: Let . . . (.) . be a possible structure for the matrix of genotypes $S_{<1,7>}$ in Table 2.1. The following is the unique parse tree that produces it:

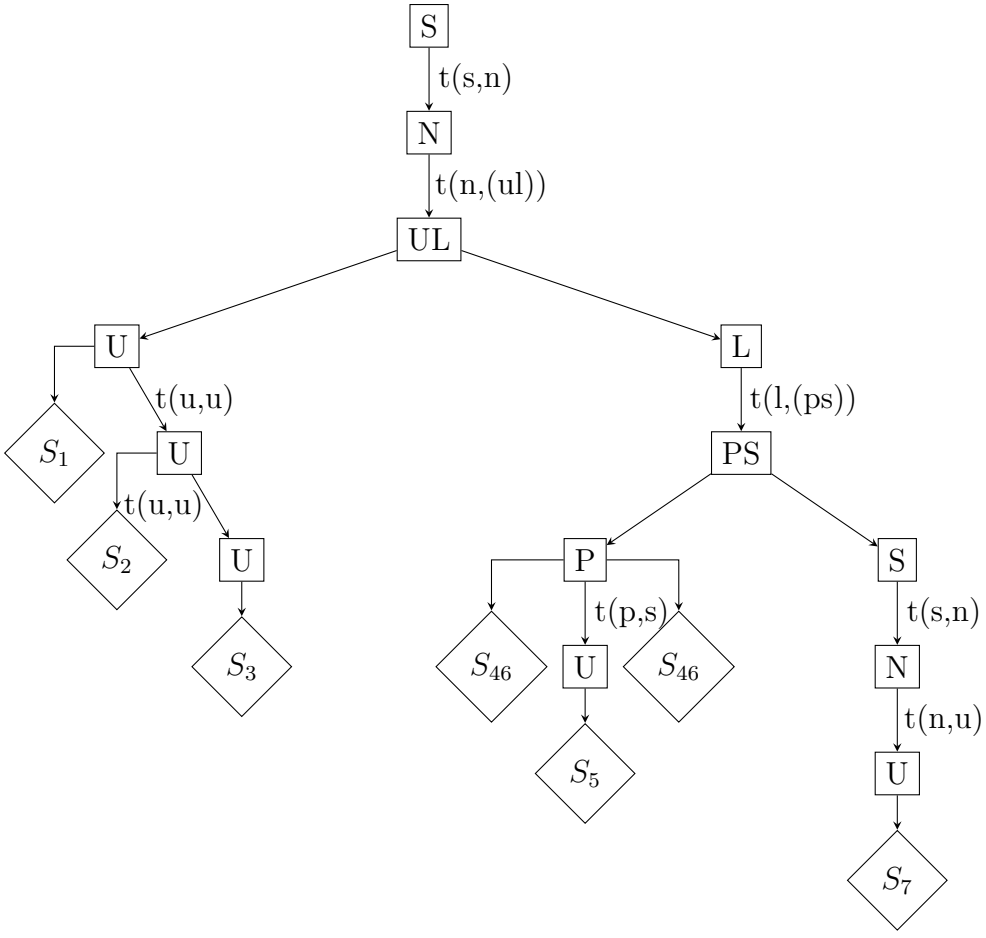


Figure 2.1: *Example Continued: The SCFG in action.* This is the unique parse tree that produces the given structure, with the non-terminal nodes are in boxes and the terminal leafs in diamonds.

2.3 Emission Parameters

The SCFG has two different types of emission distributions. The first is the distribution for the probability of the genotype counts of single SNPs emitted from the state U and the second is the distribution for the probability of the genotypes counts of associated pairs of SNPs emitted by state P.

For both types of emissions, the genotype counts are assumed to be drawn from a Multinomial distribution. All SNPs and pairs of SNPs cannot be expected to follow exactly the same model, in fact allele frequency and thus genotype frequency varies from SNP to SNP and from population to population. However, under HWE, we can expect them to be somewhat similar. By letting the parameters of the Multinomial distributions be drawn from a hierarchal model, the parameters can differ, but will all be similar. Thus, having the advantage of allowing the model for each SNP or each pair of SNPs to differ without having to greatly increase the number of parameters.

A natural assignment of a model prior is the Dirichlet distribution as it is conjugate to the Multinomial distribution. In this model, the parameters of the Multinomial are distributed according to a Dirichlet distribution where the parameters essentially pseudo-counts for the various genotypes.

The probability of the vector of genotype counts for an single SNPs i, S_i, given that it is emitted from the grammar state U, is modeled by following distribution:

$$Prob(S_i|\vec{p}) \sim Multinomial(\vec{p})$$

$$\vec{p} \sim Dirichlet(\vec{\beta})$$

where $\vec{p} = \{p_k | k = 0:2\}$ and $\vec{\beta} = \{\vec{beta}k | k = 0:2\}$.

The probability of emitting S_i from the state U is $Prob(S_i|\vec{\beta})$:

$$\begin{aligned} Prob(S_i|\vec{\beta}) &= \int Prob(S_i|\vec{p}) * Prob(\vec{p}|\vec{\beta}) d\vec{p} \\ &= \int \left(\prod_{k=0:2} p_k^{n_{k^i}} \right) \left(\prod_{k=0:2} p_k^{\beta_k - 1} \right) \left(\frac{\Gamma\left(\sum_{k=0:2} \beta_k \right)}{\prod_k \Gamma(\beta_k)} \right) d\vec{p} \\ &= \left(\frac{\Gamma\left(\sum_{k=0:2} \beta_k \right)}{\prod_k \Gamma(\beta_k)} \right) \left(\frac{\prod_k \Gamma(\beta_k + n_{k^i})}{\Gamma\left(\sum_{k=0:2} \beta_k + \sum_{k=0:2} n_{k^i l} \right)} \right) \end{aligned}$$

The probability of the matrix of genotype counts for a pair of SNPs i and j, S_{ij}, given that it is emitted from the grammar state P, is modeled by a similar distribution:

$$Prob(S_i|\vec{p}) \sim Multinomial(\vec{p})$$

$$\vec{p} \sim Dirichlet(\vec{\alpha})$$

where $\vec{p} = \{p_{kl} | k = 0:2, l = 0:2\}$ and $\vec{\alpha} = \{\vec{\alpha_{kl}} | k = 0:2, l = 0:2\}$.

The probability of emitting S_{ij} from the state P is $Prob(S_{ij}|\vec{\alpha})$:

$$
\begin{aligned}
Prob(S_{ij}|\vec{\alpha}) &= \int Prob(S_{ij}|\vec{p}) * Prob(\vec{p}|\vec{\alpha}) d\vec{p} \\
&= \int \left(\prod_{k,l=0:2} p_{kl}^{n_{k^i l^j}} \right) \left(\prod_{k,l=0:2} p_{kl}^{\alpha_{kl}-1} \right) \left(\frac{\Gamma\left(\sum_{k,l=1:3} \alpha_{kl} \right)}{\prod_{k,l} \Gamma(\alpha_{kl})} \right) d\vec{p} \\
&= \left(\frac{\Gamma\left(\sum_{k,l=0:2} \alpha_{kl} \right)}{\prod_{k,l} \Gamma(\alpha_{kl})} \right) \left(\frac{\prod_{k,l} \Gamma(\alpha_{kl} + n_{k^i l^j})}{\Gamma\left(\sum_{k,l=0:2} \alpha_{kl} + \sum_{k,l=0:2} n_{k^i l^j} \right)} \right)
\end{aligned}
$$

2.4 Summary of the SCFG

Recall, S_i is the vector of genotypes counts for SNP i, S_{ij} is the matrix of genotype counts for the pair of SNPs (i, j) and $S_{<i,j>}$ is the matrix of genotypes for the substring of SNPs from i to j. Let $W_P = \{PN, UL, U, P\}$ and $W_N = \{L, UL, U\}$ be the subsets of non-terminal states which correspond to the possible productions for the states P and N, respectively. In addition, parameters Δ_x^L and Δ_x^R are used to account for the emission of an unpaired SNP or a pair of SNPs. These parameters take the values 1 or 0 , depending on whether or not the terminal states are emitted to the right and left by state X, ie. state U emits one left terminal state and P emits one left and one right terminal. Finally, let $emit(S_{ij}|P)$ and $emit(S_i|U)$ represent the emission probabilities that follow the *Multinomial-Dirichlet* distributions with parameters $\vec{\alpha}$ and $\vec{\beta}$ as described in Section 2.3. The SCFG, $\boldsymbol{G}$ model defined previously is summarized in Table 2.2.

State X	production rules	Δ_x^L	Δ_x^R	emission	transition
P	$P \rightarrow S_{ij}W_P S_{ij}$	1	1	$emit(S_{ij}\|P)$	$t(P, W_P)$
U	$U \rightarrow S_i U$	1	0	$emit(S_i\|U)$	1
L	$L \rightarrow PS$	0	0	1	1
N	$N \rightarrow W_N$	0	0	1	$t(N, W_N)$
S	$S \rightarrow N$	0	0	1	1
PS	$PS \rightarrow P/S$	0	0	1	1
PN	$PN \rightarrow P/N$	0	0	1	1
UL	$UL \rightarrow U/L$	0	0	1	1

Table 2.2: *The Grammar* $\boldsymbol{G}$. $W_P = \{PN, UL, U, P\}$ and $W_N = \{L, UL, U\}$ are the set of possible productions for the states P and N, respectively and Δ_X^L and Δ_X^R take the values 1 or 0 , corresponding to the emission of terminal states by non-terminal state X

In Table 2.2, the following rules have been excluded: (1) $S \rightarrow E$, (2) $U \rightarrow S_i E$ and (3) $P \rightarrow S_{ij}ES_{ij}$. All three of these rules are only applicable if : (1) the sequence is empty ($S_{<i+1,i>}$), then the probability of transition is 1, (2) the sequence is of length 1 ($S_{<i,i>}$) , then the transition probability $t(U, E)$=1 and (3) the sequence is of length 2 ($S_{<i,i+1>}$), then the transition probability $t(P, E) = 1$.

2.5 The Inside Algorithm and the Multinomial Coefficients

The probability that the genotype data $S_{<1,L>}$ is generated by the SCFG model $\boldsymbol{G}$, is obtained by the *Inside Algorithm*. The *Inside Algorithm* calculates the probability of all substring of genotypes generated by the the non-terminal state space $\boldsymbol{A}$. The algorithm beings with the shortest possible substring, proceeding outwards,

capitalizing on the previous calculated probabilities.

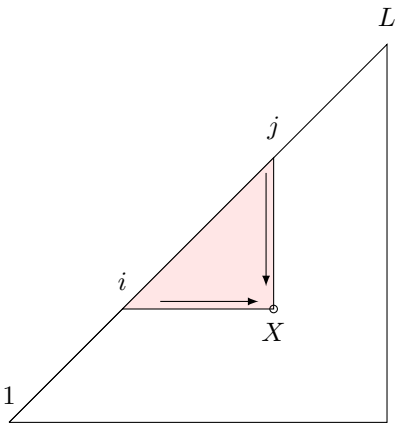


Figure 2.2: *The Inside Algorithm.* Calculates the probability of all substrings of genotypes generated by the the non-terminal state space $\boldsymbol{A}$, beginning with the shortest possible substring, proceeding outwards, capitalizing on the previous calculated probabilities.

Probability of substrings of length 1: $S_{<i,i>}$

The probability of the string $S_{<i,i>}$, given it generated state U, is the probability that state U emits the genotype counts S_i multiplied by the number of ways in which the genotype counts can be observed amongst the total number of individuals:

$$Prob(S_i|U) = \binom{N_i}{\vec{n_{k^i}}} P(S_i|\vec{\beta})$$

where $\binom{N_i}{\vec{n_{k^i}}} = \frac{N_i!}{\prod_{k=0:2} n_{k^i}!}$ and n_{k^i} is the number of times genotype k is observed for SNP i and $N_i = \sum_{k=0:2} n_{k^i}$, ie the number of individuals in the data sample. The ratio of factorials is combinatorial coefficient accounting for the number of ways in which the genotype counts n_{k^i} can be observed amongst the N_i individuals.

The probability of $S_{<i,i>}$, given it is generated by state N, is the sum of the

probability of all possible productions from N, multiplied by their corresponding transitions probabilities. While the possible productions are $N \rightarrow L|UL|U$, for a sequence of length 1, the probabilities of $S_{<i,i>}$ given L and UL are zero, because the first requires a minimum length of two and the second requires a minimum length of 3. Therefore:

$$Prob(S_i|N) = Prob(S_i|U) * t(N, U)$$

The probability of $S_{<i,i>}$, given it is generated by S, is calculated calculated similary:

$$Prob(S_i|S) = Prob(S_i|N) * t(S, N)$$

The remaining non-terminal states in $\boldsymbol{A}$ have probability zero, because the minimum required string length to be rooted in this state exceeds 1.

Probability of substrings of length 2: $S_{<i,i+1>}$

As noted in Table 2.2, states P and U are emitting states and the probabilities of the substring generated by these states is equal to the the product of the emission probability for each summary statistic and the combinatorial coefficient that counts the number of possible ways for the counts of the pair of genotypes can be observed amongst the sampled individuals.

$$
\begin{aligned}
Prob(S_{<i,i+1>}|P) &= \binom{N_{i,i+1}}{n_{k^i l^{i+1}}} P(S_{i,i+1}|\vec{\alpha}) \\
Prob(S_{<i,i+1>}|U) &= \binom{N_{i,i+1}}{n_{k^i l^{i+1}}} P(S_i|\vec{\beta}) P(S_{i+1}|\vec{\beta})
\end{aligned}
$$

where $\binom{N_{i,i+1}}{n_{k^i l^{i+1}}} = \frac{N_{i,i+1}!}{\prod_{k,l=0:2} n_{k^i l^{i+1}}!}$ and $n_{k^i l^{i+1}}$ is the number individuals with genotype k for SNP i and genotype l for SNP $i+1$ and $N_{i,i+1} = \sum_{k=0:2} \sum_{l=0:2} n_{k^i l^{i+1}}$.

The probability of generating $S_{<i,i+1>}$ from the remaining non-emitting non-terminal states, is the sum of the probability of the string over all possible states produced by the given state, multiplied by their corresponding transitions probabilities. For example, the non-empty string N can produce a string of two unpaired SNPs or string of a pair of SNPs and thus the probability $Prob(S_{<i,i+1>}|N)$ is the sum of their probabilities weighted by their corresponding transition probabilities:

$$
\begin{aligned}
Prob(S_{<i,i+1>}|PS) &= Prob(S_{<i,i+1>}|P) Prob(S_{<i+2,i+1>}|S) \\
Prob(S_{<i,i+1>}|L) &= Prob(S_{<i,i+1>}|PS) \\
Prob(S_{<i,i+1>}|N) &= Prob(S_{<i,i+1>}|L) t(N,L) + Prob(S_{<i,i+1>}|U) t(N,U) \\
Prob(S_{<i,i+1>}|S) &= Prob(S_{<i,i+1>}|N)
\end{aligned}
$$

The remaining non-terminal states have probability zero as the minimum length of the sequence generated by these states exceeds 2.

Probability of substrings of length 3: $S_{<i,i+2>}$

The probability of $S_{<i,i+2>}$, given it is generated by state P is equal to the product of the probability of emitting the genotypes of the first and last SNP as an associated pair, emitting the genotypes of the second SNP as an unpaired element and the combinatorial coefficient for the three loci. Similarly, the probability of the string rooted in state U is the coefficient multiplied by the product of the probabilities of emitting three unpaired SNPs.

$$\begin{aligned} Prob(S_{<i,i+2>}|P) &= \binom{N_{<i,i+2>}}{\vec{n}_{k^i l^{i+1} m^{i+2}}} [P(S_{i,i+2}|\vec{\alpha}) * t(P,U) * P(S_{i+1}|\vec{\beta})] \\ Prob(S_{<i,i+2>}|U) &= \binom{N_{<i,i+2>}}{\vec{n}_{k^i l^{i+1} m^{i+2}}} [P(S_i|\vec{\beta}) * P(S_{i+1}|\vec{\beta}) * P(S_{i+2}|\vec{\beta})] \end{aligned}$$

where $\binom{N_{<i,i+2>}}{\vec{n}_{k^i l^{i+1} m^{i+2}}} = \frac{N_{<i,i+2>}!}{\prod_{k,l,m=0:2} n_{k^i l^{i+1} m^{i+2}}!}$ and $n_{k^i l^{i+1} m^{i+2}}$ is the number of individuals with genotype k for SNP i, genotype l for SNP $i+1$ and genotype m for SNP $i+2$ and $N_{<i,i+2>} = \sum_{k=0:2} \sum_{l=0:2} \sum_{l=0:2} n_{k^i l^{i+1} m^{i+2}}$.

The probability of $S_{<i,i+2>}$, given it is generated by the bifurcating states UL, PS, PN, is the sum over all possible points of bifurcations of the product between the two corresponding substrings. This sum is then multiplied by the combinatorial coefficient to account for all possible arrangements of the genotype counts for these three SNPs.

For example, the probability $S_{<i,i+2>}$ given that it is rooted in state PS equal to the product of the probability of $S_{<i,i+1>}$ emitted from state P and the probability

of $S_{<i+2,i+2>}$ emitted from state U plus the product of the probability of $S_{<i,i+2>}$ emitted from state P and the probability of $S_{<i+1,i+1>}$ emitted from state U. The other bifurcations states have similar probabilities:

$$\begin{aligned}
Prob(S_{<i,i+2>}|UL) &= \binom{N_{<i,i+2>}}{n_{k^i l^{i+1}}\vec{}_{m^{i+2}}} [P(S_i|\vec{\beta}) * P(S_{i+1,i+2}|\vec{\alpha})] \\
Prob(S_{<i,i+2>}|PS) &= \binom{N_{<i,i+2>}}{n_{k^i l^{i+1}}\vec{}_{m^{i+2}}} [P(S_{i,,i+1}|\vec{\alpha}) * P(S_{i+2}|\vec{\beta}) \\
&\quad + P(S_{i,i+2}|\vec{\alpha}) * t(P,U) * P(S_{i+1}|\vec{\beta})] \\
Prob(S_{<i,i+2>}|PN) &= \binom{N_{<i,i+2>}}{n_{k^i l^{i+1}}\vec{}_{m^{i+2}}} [P(S_{i,i+1}|\vec{\alpha}) * P(S_{i+2}|\vec{\beta})]
\end{aligned}$$

For the remaining non-emitting states, the probability of the matrix of genotypes are defined as:

$$\begin{aligned}
Prob(S_{<i,i+2>}|L) &= Prob(S_{<i,i+2>}|PS) \\
Prob(S_{<i,i+2>}|N) &= Prob(S_{<i,i+2>}|L) * t(N,L) + Prob(S_{<i,i+2>}|UL) * t(N,UL) \\
&\quad + Prob(S_{<i,i+2>}|U) * t(N,U) \\
Prob(S_{<i,i+2>}|S) &= Prob(S_{<i,i+2>}|N)
\end{aligned}$$

Probability of substrings of any length: $S_{<i,j>}$

The probability of the matrix of genotypes $S_{<i,j>}$, given it is rooted in state P, is equal to the product of the probability of emitting the genotypes of SNPs i and j as an associated pair and the sum over all the possible states produced by state P of the probability of the matrix of genotypes $S_{<i+1,j-1>}$ divided by the combinatorial coefficient for the substring $< i+1, j-1 >$ string, weighted by their respective transition probabilities. This probability is multiplied by the combinatorial coefficient to account for all the ways in which the genotype counts of the substring of SNPs $< i, j >$ are possible.

$$
\begin{aligned}
Prob(S_{<i,j>}|P) \;=\; & \binom{N_{<i,j>}}{n_{k^i\vec{k^{i+1}}\ldots k^j}} P(S_{ij}|\vec{\alpha}) \\
& \times[t(P,P)Prob_{star}(S_{<i+1,j-1>}|P) + t(P,U)Prob_{star}(S_{<i+1,j-1>}|U) \\
& + t(P,UL)Prob_{star}(S_{<i+1,j-1>}|UL) \\
& + t(P,PN)Prob_{star}(S_{<i+1,j-1>}|PN)]
\end{aligned}
$$

where:

$$
\binom{N_{<i,j>}}{n_{k^i\vec{k^{i+1}}\ldots k^j}} = \frac{N_{<i,j>}!}{\prod\limits_{k^i,k^{i+1}\ldots k^j=0:2} n_{k^ik^{i+1}\ldots k^j}!}
$$

and $Prob_{star}(S_{<i,j>}|X) = Prob(S_{<i,j>}|X) \;/\; \frac{N_{<i,j>}!}{\prod\limits_{k^i,k^{i+1}\ldots k^j=0:2} n_{k^ik^{i+1}\ldots k^j}!}$.

The combinatorial coefficient of $S_{<i+1,j-1>}$ is divided out, because the combinatorial coefficient of $S_{<i,j>}$ must be multiplied in to the sum in order to account for

the possible genotype combinations for all of the SNPs in the larger substring. The probability of $S_{<i,j>}$ is based on the probability of the previously computed probability of $S_{<i+1,j-1>}$. Thus each new calculation capitalizes on the results from the previous step, for all non-terminal states.

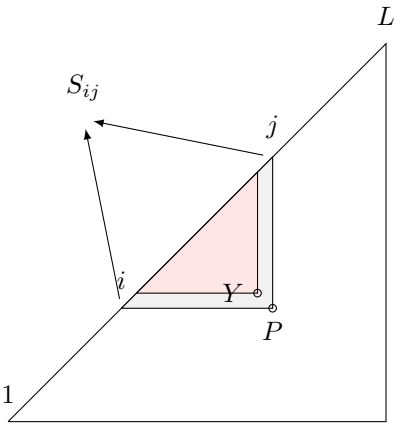


Figure 2.3: Recursion for State P. The probability $S_{<i,j>}$, given it is rooted in state P, is equal to the product of the probability of emitting paired counts $S_{i,j}$ and the probability of the matrix of genotypes $S_{<i+1,j-1>}$*generatedby*X*weightedby*t(P,X) summed over all possible X.

The probability of the matrix of genotypes of the substring of SNPs rooted in state U is the product of the probability of emitting SNP i and the probability of the matrix of genotypes $S_{<i+1,j-1>}$ given it is rooted in state U divided by the combinatorial coefficient for the substring $< i+1, j-1 >$. Again this probability is multiplied by the coefficient for the substring $< i, j >$.

$$Prob(S_{<i,j>}|U) \quad = \quad \binom{N_{<i,j>}}{n_{k^i \vec{k^{i+1}} \ldots k^j}} P(S_i|\vec{\beta}) Prob_{star}(S_{<i+1,j>}|U)$$

The probability of the string generated by one of the bifurcating states (UL, PS, PN)

is the sum over all possible points of bifurcation of the product of the substring from the left most element to the bifurcation point and the substring from element following the bifurcation point to the right most element of the string given that they are rooted in the in the left and right states of the bifurcation respectively. Each of these probabilities is divided by its respective coefficient and the sum is multiplied by the coefficient for the string $S_{<i,j>}$.

$$
\begin{aligned}
Prob(S_{<i,j>}|UL) &= \binom{N_{<i,j>}}{n_{k^i \vec{k^{i+1}} \ldots k^j}} Prob_{star}(S_{<i,i>}|U) Prob_{star}(S_{<i+1,j>}|L) \\
Prob(S_{<i,j>}|PN) &= \binom{N_{<i,j>}}{n_{k^i \vec{k^{i+1}} \ldots k^j}} Prob_{star}(S_{<i,w>}|P) Prob_{star}(S_{<w+1,j>}|N) \\
Prob(S_{<i,j>}|PS) &= \binom{N_{<i,j>}}{n_{k^i \vec{k^{i+1}} \ldots k^j}} Prob_{star}(S_{<i,w>}|P) Prob_{star}(S_{<w+1,j>}|S)
\end{aligned}
$$

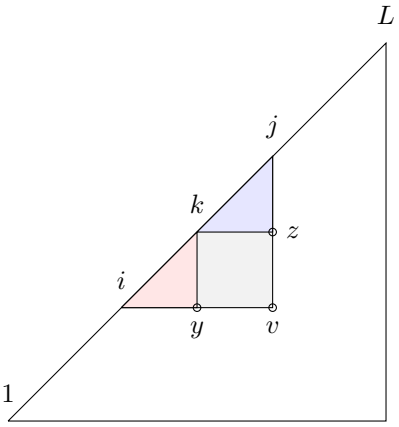


Figure 2.4: Recursion for Bifurcating States. The probability $S_{<i,j>}$, given it is rooted in a bifurcating non-terminal state v, is equal to the product of the probability of $S_{<i,k>}$ generated by state y and the probability of $S_{<k+1,j>}$ generated by z over all k, given $v \rightarrow y/z$

For the remaining non-emitting states:

$$
\begin{aligned}
Prob(S_{<i,j>}|L) &= Prob(S_{<i,j>}|PS) \\
Prob(S_{<i,j>}|N) &= Prob(S_{<i,j>}|L)t(N,L) + Prob(S_{<i,j>}|UL)t(N,UL) \\
&\quad + Prob(S_{<i,j>}|U)t(N,U) \\
Prob(S_{<i,j>}|S) &= Prob(S_{<i,j>}|N)
\end{aligned}
$$

The conditional probability of the matrix of genotypes for a sequence of SNPs $< i, j >$ being generated by a non-terminal state X given that $Y \rightarrow X$ can now be calculated easily. When Y is a non-emitting non-terminal state, the conditional probability can be calculated as in the following example for $X = L$ and $Y = N$:

$$
Prob(S_{<i,j>}, L|N) = \frac{Prob(S_{<i,j>}|L) * t(N,L)}{Prob(S_{<i,j>}|N)}
$$

When Y is an emitting non-terminal state, the conditional probability can be calculated as in the following example for $X = PS$ and $Y = P$:

$$
\begin{aligned}
Prob(S_{<i,j>}, PS|P) &= \frac{\binom{N_{<i-1,j+1>}}{n_{k^{i-1}\vec{k^i}\ldots k^{j+1}}} Prob_{star}(S_{<i,j>}|PS) * t(P,PS) * P(S_{i-1.j+1}|\vec{\alpha})}{Prob(S_{<i-1,j+1>}|P)} \\
&= \frac{\binom{N_{<i-1,j+1>}}{n_{k^{i-1}\vec{k^i}\ldots k^{j+1}}} Prob_{star}(S_{<i,j>}|PS) * t(P,PS) * P(S_{i-1.j+1}|\vec{\alpha})}{\binom{N_{<i-1,j+1>}}{n_{k^{i-1}\vec{k^i}\ldots k^{j+1}}} \sum_{y:t(P\rightarrow y)>0} (Prob_{star}(S_{<i,j>}|y) * t(P,y) * P(S_{i-1.j+1}|\vec{\alpha}))} \\
&= \frac{Prob_{star}(S_{<i,j>}|PS) * t(P,PS) * P(S_{i-1.j+1}|\vec{\alpha})}{\sum_{y:t(P\rightarrow y)>0} (Prob_{star}(S_{<i,j>}|y) * t(P,y) * P(S_{i-1.j+1}|\vec{\alpha}))}
\end{aligned}
$$

where:

$$\binom{N_{<i-1,j+1>}}{n_{k^{i-1}\vec{k^i}\ldots k^{j+1}}} = \frac{N_{<i-1,j+1>}!}{\prod_{k^{i-1},k^i\ldots k^{j+1}=0:2} n_{k^{i=1}k^i\ldots k^{j+1}}!}$$

The numerator and the denominator have the same combinatorial coefficient and thus cancel out. Therefore, the combinatorial coefficients can be dropped while calculating the probabilities of $S_{<i,j>}$ given it is generated by any non-terminal state X. The *Inside Algorithm* then recursively calculates the probabilities of all substrings over all non-terminal states. For example, if $X = P$:

$$\begin{aligned}
Prob(S_{<i,j>}|P) &\sim P(S_{ij}|\vec{\alpha})[t(P,P) * Prob(S_{<i+1,j-1>}|P) + t(P,U) * Prob(S_{<i+1,j-1>}|U) \\
&+ t(P,UL) * Prob(S_{<i+1,j-1>}|UL) + t(P,PN) * Prob(S_{<i+1,j-1>}|PN)] \\
&= P(S_{ij}|\vec{\alpha}) \sum_{(y|t(P,y)>0)} t(P,y) Prob(S_{<i+1,j-1>}|y) \\
&= emit(S_{ij}|P) \sum_{(y|t(P,y)>0)} t(P,y) Prob(S_{<i+\Delta_P^L, j-\Delta_P^R>}|y)
\end{aligned}$$

where $emit(S_{ij}|P)$ and Δ_P^L, Δ_P^R are defined in the Table 2.2.

Algorithm 1 The Inside Algorithm

Let $in(i,j,X) = Prob(S_{<i,j>}|X)$
Initializastion:
for $i = 1$ to *Length of Sequence* **do**
  $in(i+1,i,S) = 1$
end for
for $i = 1$ to L **do**
  $in(i,i,U) = emit(i,i|U)$
  for i=1 to $L-1$ **do**
    $in(i,i+1,P) = emit(i,j|P)$
  end for
end for
Recursion:
for $j = 1$ to *Length of Sequence* **do**
  for $i = j$ to 1 **do**
    for $v = P$ to S **do**

$$in(i,j,v) = \begin{cases} \sum_{k=i}^{j} in(i,k,y)in(k+1,j,z) & v = UL, PN, PS \\ emit(i,j|v) \sum_{(y|t(v,y)>0)} t(v,y)in(i+\Delta_v^L, j-\Delta_v^R, y) & otherwise \end{cases}$$

    end for
  end for
end for

where y, z are the non-terminal child states st $v \to y/z$ and Δ_x^L , Δ_x^R are defined as the number of symbols emitted to the right and left by state X as in Table 1. And furthermore:

$$emit(i,j|v) = \begin{cases} P(S_{i,j}|\vec{\alpha}) & v = P \\ P(S_i|\vec{\beta}) & v = U \\ 1 & otherwise \end{cases}$$

The final value calculated in the recursion will be $in(1,L,S)$ which is equal to the probability of matrix of genotypes for the entire sequence of SNPs given the grammar model and its associated parameters ($Prob(S_{<1,L>}|\Theta)$).
Note: The sequence of states must be transversed in the following order : $\{P, U, UL, PS, PN, L, N, S\}$ to assure that all values necessary to calculate the probability of a specific substring rooted at state v have been precomputed.

2.6 The Outside Algorithm

The Inside Algorithm is used to calculate the joint probability of the data sequence given the model. The parameters of the model are not a priori defined by the biology and thus they must be trained. In the next section, the Expectation Maximization known as the *Inside-Outisde Algorithm* will be used to learn the parameter values. In the "E" step of this algorithm, the probability of the hidden states at each position or pairs of positions is used to calculate the expected values of productions and sufficient statistics. The *Inside Algorithm* and *Outside Algorithm* are used to calculate the probability of the hidden states.

The *Outside Algorithm* calculates the probability of the matrix of genotypes for the entire string of SNPs excluding the genotypes of the string of SNPs from i to j rooted in state X ($Prob(S_{<1,L>/<i,j>}, X)$). This algorithm starts by excluding the longest possible substring, and proceeds inwards, capitalizing on the previously calculated probabilities.

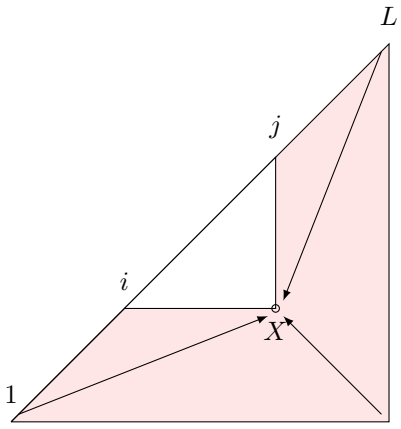


Figure 2.5: *The Outisde Algorithm.* Calculates the probability of $S_{<1,L>}$ excluding the genotypes of substring of SNPs $S_{<i,j>}$ rooted in state X, starting by excluding the longest possible substring, and proceeds inwards.

The Probability of the Empty String

The algorithm begins by calculating the probability of the empty string, i.e. the probability of excluding the genotypes of the entire string of SNPs from $< 1, L >$ rooted in the start state S. Naturally the probability of this empty string is 1 and this will be the initialization step of *The Outside Algorithm*:

$$Prob(S_{<1,L>/<1,L>}, S) = 1$$

The probability of $S_{<1,L>}$, excluding a subsequence rooted at non-terminal X from the space $\boldsymbol{A}$, is based on the productions which result in the subsequence of SNPs rooted at X. For example, state U can be produced from the states $\{N, U, P, UL\}$. If non-terminal state U generates $S_{<1,L>}$, the emitting non-terminal states U and P could not produce such a string. In addition, the bifurcating non-terminal state UL requires a string with at least two elements to the right of the string rooted in U, so again in this case this production is not valid. Thus the only production that can result in the string of SNPs $< 1, L >$ rooted in U is the rule $N \rightarrow U$. Therefore, the probability of the empty string, having excluded $S_{<1,L>}$ generated from non-terminal state U, is equal to the probability of the empty string, having excluded $S_{<1,L>}$ generated from non-terminal state N weighted by the transition probability of the production rule. Thus the probability can be expressed as follows:

$$Prob(S_{<1,L>/<1,L>}, U) = Prob(S_{<1,L>/<1,L>}, N) * t(N, U)$$

The probability of the empty sequence, having excluded $S_{<1,L>}$ generated from all other non-terminal states. can be calculated in the same fashion as above and the values are as follows:

$$\begin{aligned}
Prob(S_{<1,L>/<1,L>}, N) &= Prob(S_{<1,L>/<1,L>}, S) \\
Prob(S_{<1,L>/<1,L>}, L) &= Prob(S_{<1,L>/<1,L>}, N) * t(N, L) \\
Prob(S_{<1,L>/<1,L>}, PS) &= Prob(S_{<1,L>/<1,L>}, L) \\
Prob(S_{<1,L>/<1,L>}, UL) &= Prob(S_{<1,L>/<1,L>}, N) * t(N, UL) \\
Prob(S_{<1,L>/<1,L>}, P) &= Prob(S_{<1,L>/<1,L>}, PS) * Prob(S_{<L+1,L>}|S) \\
Prob(S_{<1,L>/<1,L>}, PN) &= 0
\end{aligned}$$

The Probability of $S_{<1,L>}$, excluding $S_{<1,L-1>}$

The only valid production rule that results in the vector of genotypes for SNPs $< 1, L-1 >$ rooted in state P is : $PS \rightarrow P/S$. Thus the probability $Prob(S_{<1,L>/<1,L-1>}, P)$ is equal to the probability of $S_{<1,L>}$, excluding the sequence of SNPs from 1 to L rooted in PS, times the probability of the right hand substring of the bifurcation $S_{<L-1,L>}$ generated by S. This latter probability is obtained from *The Inside Algorithm*. The probability can be expressed as follows:

$$Prob(S_{<1,L>/<1,L-1>}, P) = \frac{N_L!}{\prod_{k=0:2} n_k^L!}[Prob(S_{<1,L>/<1,L>}, PS) Prob(S_{<L,L>}|S)]$$

Note, both $Prob(S_{<1,L>/<1,L>}, PS)$ and $Prob(S_{<L,L>}|S)$ have been previously calculated in the previous step and the *Inside Algorithm*, respectively.

For the remaining non-terminal states, the probability $Prob(S_{<1,L>/<1,L-1>}, X)$ $= 0$ due to the fact that there is no production rule that will make the probability defined.

The Probability of $S_{<1,L>}$, excluding $S_{<1,L-2>}$

Similarly the probability of the last two elements of the sequence, having excluded $S_{<1,L-2>}$ generated by the non-terminal state P, can be calculated as follows:

$$Prob(S_{<1,L>/<1,L-2>}, P) \quad = [Prob(S_{<1,L>/<1,L>}, PS) Prob(S_{L-1,L}|S)]$$

The only valid production rule that results in $S_{<1,L-2>}$ rooted in state U is : $UL \rightarrow U/L$. Thus the probability $Prob(S_{<1,L>/<1,L-2>,U}$, is equal to the probability of $S_{<1,L>}$, excluding $S_{<1,L>}$ rooted in PS, times the probability of the right hand substring $S_{<L-2,L>}$ generated by L:

$$Prob(S_{<1,L>/<1,L-2>}, U) = [Prob(S_{<1,L>/<1,L>}, UL) Prob(S_{L-1,L}|L)]$$

Again, both $Prob(S_{<1,L>/<1,L>}, UL)$ and $Prob(S_{L-1,L}|L)$ have been previously calculated in the previous step and the *Inside Algorithm*, respectively.

There are no valid production rules that would produce $S_{<1,L-2>}$rooted in the remaining non-terminals and thus the $Prob(S_{<1,L>/<1,L-2>}, X) = 0$ for all remaining non-terminals in $\boldsymbol{A}$.

The Probability of $S_{<1,L>}$, excluding $S_{<i,j>}$

The probability of the genotypes for the entire sequence $S_{<1,L>}$, excluding the genotypes of any possible substring $S_{<i,j>}$ rooted in any non-terminal state X, is equal to the sum of the probabilities over all possible productions of $S_{<i,j>}$ rooted in state X.

For non-terminal S, the only rule that produces it is $PS \rightarrow P/S$ and the state PS can generate all sequences $S_{<k,j>}$ for $k = 1 : i-2$ to allow for P to generate at least one paired terminal state. Thus the $Prob(S_{<1,L>/<i,j>}, S)$ is equal to the sum of the probability $Prob(S_{<1,L>/<k,j>}, PS)$ over all possible k multiplied by the probability of the sequence $S_{<k,p>}$ given that it is generated by state P. The probability can be expressed as follows:

$$Prob(S_{<1,L>/<i,j>}, S) \quad = \quad \sum_{k=1:i-2} Prob(S_{<1,L>/<k,j>}, PS) Prob(S_{<k,i-1>}|P)$$

Along the same lines, for non-terminal state N, the productions rules that produce it are $PN \rightarrow P/N$ and $S \rightarrow N$. Thus the probability can be expressed as follows:

$$Prob(S_{<1,L>/<i,j>}, N) = \sum_{k=1:i-2} Prob(S_{<1,L>/<k,j>}, PN) Prob(S_{<k,i-1>}|P) + Prob(S_{<1,L>/<i,j>}, S)$$

For non-terminal state L, the productions rules that produce it are $UL \rightarrow U/L$ and $N \rightarrow L$. Thus the probability can be expressed as follows:

$$Prob(S_{<1,L>/<i,j>}, L) = \sum_{k=1:i-1} Prob(S_{<1,L>/<k,j>}, UL) Prob(S_{<k,i-1>}|U) + Prob(S_{<1,L>/<i,j>}, N) * t(N, L)$$

For non-terminal state PN, the production rule that produces it is $P \rightarrow PN$. Since P is an emitting state, P generates the sequence $S_{<i-1,J+1>}$. It emits pair terminal $S_{i-1.j+1}$ and transitions to PN. Thus the probability can be expressed as follows:

$$Prob(S_{<1,L>/<i,j>}, PN) = Prob(S_{<1,L>/<i-1,j+1>}, P) * t(P, PN) * P(S_{i-1,j+1}|\vec{\alpha})$$

For non-terminal state PS, the production rule that produces it is $L \rightarrow PS$. This is not an emitting state and the probability of transition is 1. Thus the probability

can be expressed as follows:

$$Prob(S_{<1,L>/<i,j>}, PS) \quad = \quad Prob(S_{<1,L>/<i,j>}, L)$$

For non-terminal state UL, the production rules that produce are: $P \rightarrow UL$ and $N \rightarrow UL$. As above, since P is an emitting state, it generates $S_{<i-1,j+1>}$, emits pair terminal $S_{i-1.j+1}$ and transitions to UL. Thus the probability can be expressed as follows:

$$\begin{aligned} Prob(S_{<1,L>/<i,j>}, UL) \quad = \quad & Prob(S_{<1,L>/<i-1,j+1>}, P) * t(P, UL) * P(S_{i-1,j+1}|\vec{\alpha}) \\ & + Prob(S_{<1,L>/<i,,j>}, N) * t(N, UL) \end{aligned}$$

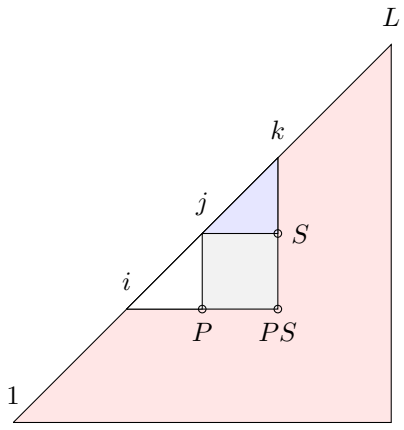
For non-terminal state U, the production rules that produce are: $U \rightarrow U$, $P \rightarrow U$, $N \rightarrow U$ and $UL \rightarrow U/L$. Following the same reasoning above for emitting and bifurcating states, the probability can be expressed as follows:

$$
\begin{aligned}
Prob(S_{<1,L>/<i,j>}, U) \;=\; & Prob(S_{<1,L>/<i-1,j>}, U) * t(U,U) * P(S_{i-1}|\vec{\beta}) \\
& + Prob(S_{<1,L>/<i-1,j+1>}, P) * t(P,U) * P(S_{i-1,j+1}|\vec{\alpha}) \\
& + Prob(S_{<1,L>/<i,j>}, N) * t(N,U) \\
& + \sum_{k=j:L} Prob(S_{<1,L>/<i,k>}, UL) Prob(S_{<j+1,k>}|L)
\end{aligned}
$$

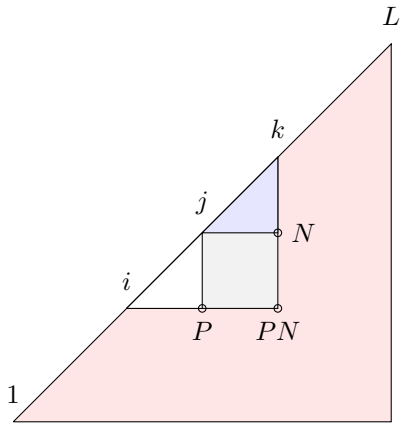
Finally, for non-terminal state P, the production rules that produce are: $P \rightarrow P$, $PS \rightarrow P/S$, $PN \rightarrow P/N$. Following the same reasoning above for emitting and bifurcating states, the probability can be expressed as follows:

$$
\begin{aligned}
Prob(S_{<1,L>/<i,j>}, P) \;=\; & Prob_{star}(S_{<1,L>/<i-1,j+1>}, P) * t(P,P) * P(S_{i-1,j+1}|\vec{\alpha}) \\
& + \sum_{k=j:L} [Prob(S_{<1,L>/<i,k>}, PS) Prob(S_{<j+1,,k>}|S) \\
& + \sum_{k=j:L} [Prob(S_{<1,L>/<i,k>}, PN) Prob(S_{<j+1,k>}|N)]
\end{aligned}
$$

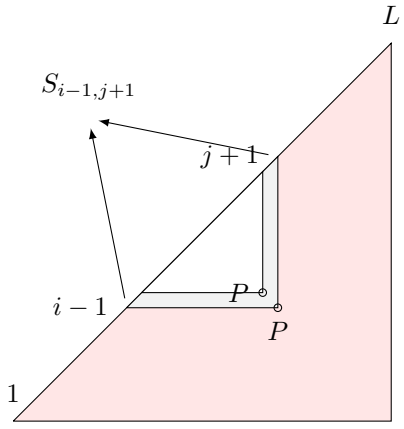
For all i and j, the probability calculations depend on previously calculated probabilities of shorter subsequences and the previously calculated probabilities from the *Inside Algorithm.*



(a) $PS \rightarrow P/S$



(b) $PN \rightarrow P/N$



(c) $P \rightarrow P$

Figure 2.6: Outside Recursion for State P. The production rules that produce are: $P \rightarrow P$, $PS \rightarrow P/S$, $PN \rightarrow P/N$ and the probability of $S_{<1,L>}$ excluding $S_{<i,j>}$ rooted in state P is the sum over the probabilities associated to these three productions.

Algorithm 2 The Outside Algorithm

Let $Prob(S_{<1,L>/<i,j>}, X) = out(i,j,X)$ and $in(i,j,X) = Prob(S_{<i,j>}|X)$
Initialization:
$out(1, L, S) = 1$
Recursion:
for $i = 1$ to L+1 **do**
  for $j = L$ to $i-1$ **do**
    for $v = S$ to P **do**

$$out(i,j,v) = \begin{cases} \sum_{k=1}^{i} out(k,j,y) in(k,i-1,z) & v = L, N, S \\ \sum_{k=j}^{L} out(i,k,y) in(j+1,k,z) & v = U, P \\ emit(i-\Delta_y^L, j+\Delta_y^R|y) \sum_{(y|t(y,v)>0)} t(y,v) out(i-\Delta_y^L, j+\Delta_y^R, y) & all\ v \end{cases}$$

    end for
  end for
end for
where y, z are the non-terminal states st $y \to z/v$ for $v = \{L, N, S\}$, $y \to v/z$ for $v = \{U, P\}$ and Δ_x^L ,Δ_x^R are defined as the number of symbols emitted to the right and left by state X as in Table 1. And furthermore:

$$emit(i,j|v) = \begin{cases} P(S_{i,j}|\vec{\alpha}) & v = P \\ P(S_i|\vec{\beta}) & v = U \\ 1 & otherwise \end{cases}$$

2.7 Estimating The Model Parameters: The Inside-Outside Algorithm

For the polymorphic genotype data, there are two sets of unknowns: the model parameters θ and the hidden data (the location of the independent and associated loci). The Expectation Maximization method known as the *Inside-Outside Algorithm*

is used to iteratively train the model parameters.

In the "E" expectation step, the probability of the hidden data is used to find the expected values of the productions and genotype counts. The expected number times non-terminal state v is visited is equal to the sum of the condition probability that state v generates $S_{<i,j>}$ for all i and j:

$$\begin{aligned} EC(v) &= \sum_{i=1:L}(\sum_{j=i:L}\frac{Prob(S_{<i,j>}|v)Prob(S_{<1,L>/<i,j>},v)}{Prob(S_{<1,L>}|\theta)}) \\ &= \sum_{i=1:L}\sum_{j=i:L}\frac{in(i,j,v)out(i,j,v)}{in(i,L,S)} \end{aligned}$$

for all non-terminal states $v \in \boldsymbol{A}$.

Similarly, the expected number of times non-terminal state v produces non-terminal state y, for all v and y $\boldsymbol{A}$ is calculated as follows:

$$\begin{aligned} EC(v,y) &= \sum_{y:(t(v,y)>0)}\sum_{i=1:L}\sum_{j=i:L}\frac{Prob(S_{<i,L>/<i,j>},v)Prob(S_{<i+\Delta_v^L,j-\Delta_v^R)>}|y)t(v,y)Prob(S_{(ij)}|v)}{Prob(S_{<i,L>}|\theta)} \\ &= \sum_{y:(t(v,y)>0)}\sum_{i=1:L}\sum_{j=i:L}\frac{out(i,j,v)emit(i,j,v)t(v,y)in(i+\Delta_v^L,j-\Delta_v^R,y)}{in(i,L,S)} \end{aligned}$$

where $Prob(S_{(ij)}|v) = [Prob(S_i|v)I(v = U) + Prob(S_{ij}|v)I(v = P) + I(v = \{S,N,L,UL,PS,PN\})]$.

In the "M" step, these expected values are used to calculate the maximum like-

lihood estimates of the model parameters. The transition probabilities from a given non-terminal state v follow a multinomial distribution, thus the maximum likelihood estimate of these parameters is the frequency of the expected values:

$$\hat{t}(v,y) = \frac{EC(v,y)}{EC(v)}$$

for all v and y in the non-terminal state space $\boldsymbol{A}$.

For emission distributions, there are two sets of parameters that must be estimated, $\vec{\alpha}$ and $\vec{\beta}$. The former is the parameter vector for the nine category Multinomial-Dirichlet distribution modeling the emission of the genotype counts for a pair of SNPS from state P and the latter is the parameter vector for the three category Multinomial-Dirichlet modeling the emission of genotype counts for a single SNP from the state U.

For the former, it is necessary to calculate the expected genotype counts for all possible pairs of SNPs given that the pair of SNPs is emitted on a production from state P using the probabilities obtained from *The Inside Algorithm* and *The Outside Algorithm*. The only way for S_{ij} to be emitted from a production from state P is for $S_{<i,j>}$ to be rooted in state P, thus:

$$\begin{aligned} EC_P(S_{ij}) &= \frac{Prob(S_{<i,j>}|P) * Prob(S_{<1,L>/<i,j>}, P)}{Prob(S_{<1,L>}|S)} * S_{ij} \\ &= \frac{in(i,j,P) * out(i,j,P)}{in(1,L,S)} * S_{ij} \end{aligned}$$

where the the numerator is the joint probability of the entire data sequence $S_{<1,L>}$ given the model and $S_{<i,j>}$ being rooted in state P. The numerator is divided by the the probability of $S_{<1,L>}$ given the model. This results in conditional probability of S_{ij} being emitted from state P.

For the later, it is necessary to calculate the expected genotype counts for all SNPs given the SNP is emitted from a production from state U. Unlike in the previous case, there are many ways to emit S_i from U. When SNP i is the left most element of any substring of SNPs rooted in state U, the genotypes of S_i will be emitted on the production, thus it is necessary to sum over strings $< i, j >$ of SNPs for all $j = i : L$ as follows:

$$\begin{aligned} EC_P(S_i) &= \sum_{j=i:L} \frac{Prob(S_{<i,L>}|U) * Prob(S_{<1,L>/<i,j>}, U)}{Prob(S_{<1,L>}|S)} * S_i \\ &= \sum_{j=i:L} \frac{in(i,j,U) * out(i,j,U)}{in(1,L,S)} * S_i \end{aligned}$$

Having calculated the expected genotype counts for each emission state (U and P), the M step will require finding the parameters which maximize the likelihood of the expected data for each of these two state. The likelihood of the expected genotype counts given the unpaired model is defined as:

$$
\begin{aligned}
Prob(D|\vec{\alpha}) &= \prod_i Prob(S_i|\vec{\alpha}) \\
&= \prod_i \left(\frac{\Gamma\left(\sum_{k=1:3} \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} * \frac{\prod_k \Gamma(\alpha_k + n_{k^i})}{\Gamma\left(\sum_{k=1:3} \alpha_k + \sum_{k=1:3} n_{k^i}\right)} \right)
\end{aligned}
$$

where $\vec{\alpha} = \{\alpha_1, \alpha_2, \alpha_3\}$. To find $\hat{\alpha}$ that maximizes this likelihood of the data we will follow the algorithm by Thomas Minka (2000) in a simplified Newton iteration is used by taking the Hessian of the log-likelihood:

$$
\begin{aligned}
\frac{dlogP(D|\vec{\alpha})}{d\alpha_k^2} &= \sum_i \Psi'(\sum \alpha_k) - \Psi'(N_i + \sum \alpha_k) + \Psi'(n_{k^i} + \alpha_k) - \Psi'(\alpha_k) \\
\frac{dlogP(D|\vec{\alpha})}{d\alpha_k \alpha_j} &= \sum_i \Psi'(\sum \alpha_k) - \Psi'(N_i + \sum \alpha_k)
\end{aligned}
$$

and writing the hessian as a matrix st:

$$
\begin{aligned}
\boldsymbol{H} &= \boldsymbol{Q} + \mathbf{1}\mathbf{1}^{\boldsymbol{T}} z \\
z &= \sum_i \Psi'(\sum \alpha_k) - \Psi'(N_i + \sum \alpha_k) \\
q_{jk} &= \delta(j-k) \sum_i \Psi'(n_{k^i} + \alpha_k) - \Psi'(\alpha_k)
\end{aligned}
$$

For one Newton Iteration:

$$\vec{\alpha_{new}} = \vec{\alpha_{old}} - \boldsymbol{H^{-1}g}$$

where $\boldsymbol{g}$ is the gradient of $logP(D|\vec{\alpha})$. Minka is able to compute $\boldsymbol{H^{-1}}$ without storing or inverting the hessian but rather it is computed via $\boldsymbol{Q^{-1}}$and z as defined above. Details on this computation as well as the stopping rule can be found in the 2000 paper by Minka. A similar computation is done for the expected counts data for the associated state P using the nine category $\vec{\alpha} = \{\alpha_{kl}, k, l = 1:3\}$.

Iteration between the expectation and maximization steps above continues until joint probability of the data given estimated model parameters converges.

2.8 The Backsampling Algorithm

While it is possible to obtain the the most probable derivation/LD structure of the data using a variation of *The Inside Algorithm*, this structure does not always give the best representation of the space of possible LD structures. The centroid has been shown to provide a better representation of the ensemble of structures. It has been shown that the it is possible to calculate the centroid by taking a representative sample of the ensemble and finding all associations that appear greater than 50 percent of the time. In addition, this model has imposed constraints that are not necessarily biologically driven. Sampling from the ensemble will allow for the identification of variations in the structure that would otherwise not found from the *The Inside Algorithm* and *The Outside Algorithm*.

With *The Inside Algorithm*, the probability of the matrix of genotypes for the

given sequence of SNPs, $S_{<1,L>}$ being generated by model and its parameters is calculated. With this partition function completely defined, exact back-samples from the posterior probability distribution can be obtained.

To back-sample a parse tree, the first state in the sample is the non-terminal start state S, which is the root of $S_{<1,L>}$. From the start state S a production is sampled in proportion to the posterior probability distribution that can be obtained from the *Inside Algorithm*. From the start state S, the only production is $S \rightarrow N$. Thus, with probability 1, $S_{<1,L>}$ is generated by non-terminal state N in the sampled parse tree. The possible production rules from non-terminal state N are $N \rightarrow L|\,UL|\,U$ and are then sampled according to their posterior probabilities.

The posterior probability of non-terminal state L being the root of the sequence $S_{<1,L>}$ given that non-terminal state N is the root of the sequence $S_{<1,L>}$ is equal to the product of the probability of $S_{<1,L>}$ given that it is generated by L and the probability of L given N, divided by the the marginal probability of the data sequence $S_{<1,L>}$ given it is generated by state N.

$$\begin{aligned} Prob(L|S_{<1,L>}, N) &= \frac{Prob(S_{<1,L>}|L) * Prob(L|N)}{Prob(S_{<1,L>}|N)} \\ &= \frac{in(1,L,L) * t(N,L)}{in(1,L,N)} \end{aligned}$$

The same calculation is computed for the other states UL and U:

$$
\begin{aligned}
Prob(UL|S_{<1,L>},N) &= \frac{in(1,L,UL)*t(N,UL)}{in(1,L,N)} \\
Prob(U|S_{<1,L>},N) &= \frac{in(1,L,U)*t(N,U)}{in(1,L,N)}
\end{aligned}
$$

Given these probabilities, if the production rule $N \rightarrow UL$ is sampled, UL will be the root of $S_{<1,L>}$ in the sampled parse tree. Non-terminal state UL is one of the three non-terminal bifurcation states. With a bifurcation state, a bifurcation point k must be sampled such that the matrix $S_{<1,L>}$ generated by state UL produces two subsequence: $S_{<1,k>}$ generated by state U and $S_{<k+1,L>}$ generated by state L. The bifurcation point k is sampled in proportion its posterior probability distribution. For all $k = \{1 : L-1\}$, the posterior probability of k given that $S_{<1,L>}$ is rooted in state UL is the product of the probabilities of the two subsequences divided by the marginal probability:

$$
Prob(k|S_{<1,L>}UL) = \frac{Prob(S_{<1,k>}|U) * Prob(S_{<k+1,L>}|L)}{Prob(S_{<1,L>}|UL)}
$$

$S_{<1,k>}$ is now rooted in U and $S_{<k+1,L>}$ is now rooted in L in the sampled parse. The productions from each of these roots will be sampled simultaneously. Sampling is complete when for each bifurcated subsequence of SNPs, the position of the end of the sequence is less than the beginning.

Algorithm 3 The Backsampling Algorithm

function BACKSAMPLE($SM \leftarrow begin,\ end,\ current\_state,\ SM$)
  if $begin < end$ and $current\_state = S, N, L, U, P$ **then**
    $r = generate\_random(0, 1)$
    for $new\_state = P : S$ **do**
      $cumprob \leftarrow cumprob + Prob(new\_state|seq(begin, end), current\_state)$
      if $r < cumprob$ **then**
        $SM(begin, end, new\_state) \leftarrow 1$
        $begin \leftarrow begin + delta(left\_emission, current\_state)$
        $end \leftarrow end + delta(right\_emission, current\_state)$
        $current\_state \leftarrow new\_state$
      end if
    end for
  else if $begin < end$ and $current\_state = PN, PS, UL$ **then**
    $r = generate\_random(0, 1)$
    for $bif\_point = begin : end$ **do**
      $cumprob < cumprob + Prob(bif\_point|seq(begin, end), current\_state)$
      if $r < cumprob$ **then**
        $SM(begin, bif\_point, left\_state) \leftarrow 1$
        $SM(bif\_point + 1, end, right\_state) \leftarrow 1$
        $SM \leftarrow backsample(begin, bif\_point, left\_state, SM)$
        $SM \leftarrow backsample(bif\_point + 1, end, right\_state, SM)$
      end if
    end for
  end if
end function

Parse trees are sampled using the above function initialized with $begin = 1$, $end = L$ and $current\_state = S$. For each sampled tree, the function will return the $L \times L \times size(\boldsymbol{A})$ matrix SM. A representative sample of the ensemble of derivations of the data $S_{<1,L>}$ can be obtained by repeated back-sampling. The probability of an association between two SNPs i and j can be approximated by adding up the occurrences of $S_{<i,j>}$ rooted at state P in the sampled trees.

Chapter Three

Implementation and Results

3.1 The SCFG and Simulated Genotype Data

The SCFG model, as discussed in Chapter 1, is both a language recognizer and a language generator. To access the model's sensitivity and specificity to infer jointly occurring interactions between SNPs, the grammar was used in its generative capacity to simulate genotype data.

3.1.1 Simulation of Data

To simulate populations with different proportions of unpaired/paired SNPs as well as different variance in the distribution of genotypes/association, multiple data sets were generated from the model with varying combinations of transition parameters and emission parameters, θ. For a specified θ, a parse tree, T_θ, was generated, such that the sequence of non-terminal states, corresponds to a sequence of SNPs of length $L = 300$ and rooted at the start state S. Given the parse tree T_θ, for each of 1000 individuals, the genotypes of all unpaired SNP and all pairs of SNPs were simulated according to the specified emission distributions. The summary statistics S_i and S_{ij} were computed for the matrix of simulated genotypes.

3.1.2 Sensitivity, Specificity and PPV

Five computer simulated data sets were generated. The first data set was generated from the model with null parameters θ_0, i.e. the probability of all production rules involving state P was set at zero. The genotypes for the 300 unpaired SNPs were generated for 1000 individuals under the assumption that all individuals shared

similar ancestry and were unaffected by population stratification. The production rules involving state P were all correctly estimated as having probability zero and thus no spurious associations were inferred. For the remaining 4 simulations, data sets 1 and 2 were generated using the same parameters vectors and the 3rd and 4th was generated from parameter vectors allowing for more variation in the genotypes counts for both paired and unpaired snps.

The accuracy of the posterior predictive inference of the SCFG model was accessed by implementing *The Inside-Outside Algorithm* on the Simulated Data sets. 1000 parsed structures were sampled from the joint distribution calculated by the *Inside Algorithm* and the representative centroid was found for each data set. The Positive Predictive Value provides a quantitative measure of the predictive accuracy of the model and is calculated as the number of true positives divided by the number of positive calls. The PPV was obtained by a comparison of the simulated epistatic associations and the associations predicted in the ensemble centroid. Also of interest is the sensitivity and the specificity, reflecting the percentage of occurring associations the model is able to detect and the percentage of unpaired loci the model is able to detect, respectively.

Table 3.1 contains the sensitivity and specificity measurements and the PPV from all five data sets. True positives (TP) are the pairs which were correctly identified and True negatives (TN) are the unpaired SNPs that were correctly identified. Where as, the False Positives (FP) are pairs of SNPs incorrectly identified and False Negatives (FN) are since SNPs that were incorrectly identified. The average PPV, which is calculated as TP/(TP+FP), from the 4 non-null simulations is .97. In addition the average sensitivity, which is calculated as TP/(TP+FN), is .64 and the average specificity, which is calculated as TN/(FP+TN) is .99. From the four data sets, it appears that there is a large variance in the sensitivity.

Simulation	Number of Simulated Pairs	Sensitivity		Specificity		PPV	
Null	0	-	-	500/500	1.0	-	-
1	50	22/50	.44	398/400	.99	22/23	.96
2	42	40/42	.95	412/416	.99	40/42	.95
3	36	18/36	.50	448/448	1.0	18/18	1.0
4	116	78/116	.67	164/168	.98	78/80	.97

Table 3.1: *Computer Generated Simulations.* Four simulated data sets were generated from varying model parameters. The sensitivity was calculated as TP/(TP+FN), the specificity was calculated as TN/(FP+TN), and the PPV was calculated as TP/(TP+FP).

The results from the simulated data indicate, from the high PPV, that it is highly unlikely to infer false positive associations. In addition, while it would be ideal for both the sensitivity and specificity to be close to 1, the results show that we are able to consistently detect more than 50 percent of all associations and an even higher percentage of all statistically independent loci.

3.2 The SCFG and The 1000 Genome Project

3.2.1 Parameter Estimation with Large Data

The 1000 genome project was developed as a resource for the investigation of genetic contribution to disease by characterizing human genetic variation across the geographic and functional spectrum. This project resulted in 1092 sampled individuals from 14 populations from Europe, East Asia, Africa and the Americas. Through a variety of algorithms, the project produced a validated HAPMAP of 38 million SNPs, 1.4 million short indels, and 14000 larger deletions.

With a large amount of SNP data at hand, and the assumption that association

does not exist between loci on different chromosomes, to train the model parameters (emission and transition), 6 random chromosomal segments were sampled in proportion to the length of each chromosome, extending 40kb. For each chromosomal segment, SNPs with minimum allele frequency of $< .05$ were filtered out. *The Inside and Outside Algorithms* were applied to each chromosomal segment independently and then all expected values were pooled for the maximum likelihood estimation of the parameters. The procedure was iterated until convergence of the sum of the likelihood over all segments. This process was repeated 4 times. The estimated parameters, both emission and transition, for the 4 replications appear below in Table 3.2. The average of the parameters over the four replications, as well as the associated standard deviations appear in table.

Repeat	Total SNPs	EP:U	EP:P	TP:N	TP:P
1	726	9.5719 1.7467 0.3895	3.6302 0.1766 0.0408 0.1648 1.7774 0.1092 0.0356 0.1000 0.5885	0.8378 0.0515 0.1107	0.4574 0.1297 0.1865 0.2264
2	595	12.9045 1.8707 0.3212	3.6997 0.2021 0.0658 0.1925 1.0908 0.0965 0.0468 0.0787 0.3449	0.7779 0.2221 0.0000	0.3333 0.3332 0.1668 0.1667
3	710	11.0160 1.7831 0.3612	3.7458 0.1631 0.0449 0.1850 1.3453 0.0960 0.0513 0.0993 0.4485	0.6888 0.1498 0.1614	0.4270 0.1807 0.2489 0.1435
4	831	8.0068 1.5000 0.3573	3.0776 0.1822 0.0517 0.1740 1.1743 0.0929 0.0491 0.0835 0.3651	0.5238 0.2381 0.2381	0.3340 0.1989 0.2673 0.1998
Avg	-	10.3748 1.7251 0.3573	3.5383 0.1810 0.0508 0.1791 1.3469 0.0986 0.0457 0.0904 0.4367	0.7070 0.1654 0.1276	0.3879 0.2106 0.2174 0.1841
Std	-	2.0860 0.1588 0.0280	0.0966 0.0162 0.0109 0.0122 0.3059 0.0072 0.0070 0.0109 0.1102	0.1367 0.0851 0.0999	0.0639 0.0868 0.0483 0.0365

Table 3.2: *1000 Genomes Training Data Repeats.* The trained parameters for the four training repeats appear in the table. Starting from the third column, the table contains the emission parameters from state U, the emission parameters from the state P, the transition parameters for $N \rightarrow L|UL|U$ respectively and the transition parameters for $P \rightarrow PN|UL|U|P$, respectively.

3.2.2 Association Inference for Multiple Genomic Regions

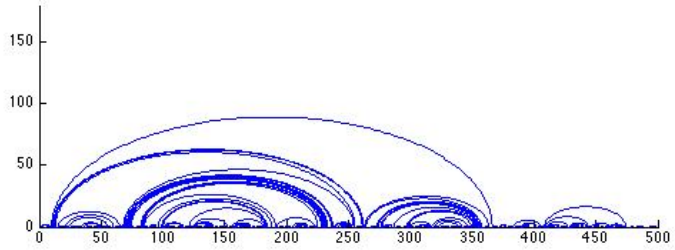
The averages of the trained parameter repeats were used as θ in the grammar model for the following inference. Three longer genomic regions were selected at random from chromosome 2, 4 and 18 with lengths of 600k, 600k, and 1mb respectively. To obtain the probability of the each genomic segment given the model $\boldsymbol{G}$ and parameters θ, *The Inside Algorithm* was implemented on each segment, independently. 1000

parse trees were sampled according to the joint probability distribution for each genomic region and the centroid was determined for each region, to find the parse that maximizes the posterior marginal sum [57]. The genomic position of the sampled segments, as well as the number of paired associations predicted in the centroid, the maximal length of the paired SNPs and the number of associated pairs of genomic distance greater than 10kb can be found in Table 3.3.

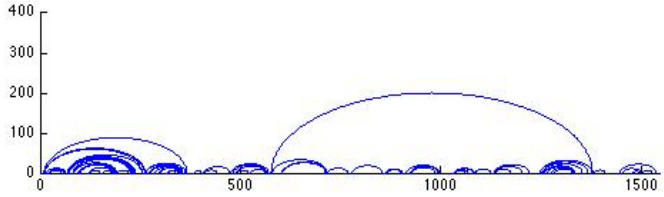
Chr	AF Start Position	AF End Position	Number of SNPs	Paired SNPS in Centroid	Max Distance between loci	Number of Pairs $> 10kb$
2	102319730	102918601	1545	632	237k	78
4	104213866	104813402	1940	784	121k	80
18	48742255	49742142	3152	1270	173k	133

Table 3.3: *The 1000 Genomes Project Centroid Analysis.* This table contains the start and end positions of the genomic segments, as well as the number of paired associations appearing in the centroid, the maximal distance between paired SNPs and the number of associated pairs of genomic distance greater than 10k.

Below is a pictorial representation of the associated pairs predicted in the centroid. The first 500 SNPs from Chr2:102319730 are depicted in the top and the entire genomic region of Chr2:102319730 is depicted in the bottom of in Figure 3.1. Both figures are the commonly used ellipse diagram, with the height of the ellipse proportional to the number of SNPs in between the endpoints of the pair. Similar plots for the genomic segment Chr4: 104213866, both in its entirety and for the first 500 SNPs of the segment are located in Appendix A.3.



(a)



(b)

Figure 3.1: *Chromosome 2:102319730 Predicted Pairs of the Centroid.* The figure is the common ellipse diagram with the height of the ellipse in proportion to the number of SNPs in between the endpoints of the pair (a) associated pairs for the first 500 SNPs in the genomics region (b) the associated pairs of genomic distance greater than 10k for the entire genomic region

From the centroids of these sampled genomic regions, there is evidence that there exists long range linkage disequilibrium at distances much greater than historically

believed, with about 10 percent of all high frequency associated SNPs being of length of at least 10kb. From these pictorial representations, it appears that the majority of the genome can be broken into 'blocks' of linkage. However, the model has also predicted some distant associated pairs whose endpoints encompass these smaller blocks, for example the pair of SNPs at the 577th and 1379th loci. The existence of such an associated pair, with this pair occurring in more than 80 percent of the representative sample, provides evidence that the regions locally linked to these loci may not be independent. Thus, the model predicts LD over distant non-contiguous regions, which may have not otherwise been predicted via blocking procedures that define blocks based on the local decay of LD and/or low allelic diversity.

Due to the constraints of the SCFG model, in one single structure, including the representative centroid, there are no crossing arcs and each loci is paired to at most one other loci. Thus the single centroid of the representative sample alone does not provide insight into more complex interactions, such as interactions that occur between more than two genomic regions. Insight into these types of associations, can be obtained from the the representative sample of the ensemble of all possible structures. Figure 3.2 contains all the pairs that were predicted in the representative back sample, colored by their frequency in the sampled structures. In the studied genomic region of Chromosome 2, while they do not appear in the centroid structure, the association between SNPs at loci 577 and 857 and association between SNPs at loci 858 and 1379 are predicted in the representative sample. These associations suggest the existence of higher order interactions, with the genetically linked regions around the SNP at the 577th and the SNP at the 1379th loci interacting with each other and a possible third region around the 857/858th loci. Thus the SCFG model for inference on pairs of pairs of pairs of associations, as well as the representative sample of structures from the posterior ensemble can be seen as projection of the

high dimensional space that lacks these constraints into a computationally tractable space with these constraints.

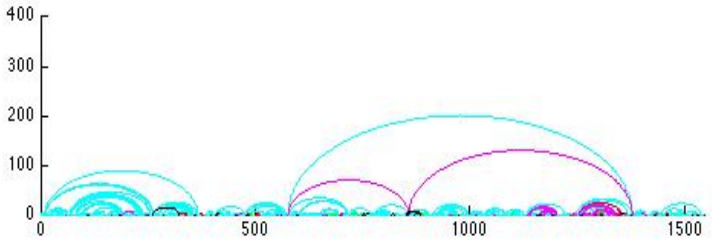


Figure 3.2: *Chromosome 2:102319730 Predicted Pairs from Back-Sampling.* 500 structures were sampled from the ensemble of all permitted pairwise association structures. The height of the ellipse corresponds to the number of SNPs between the loci and the color corresponds to the frequency of the pair in the representative sample. Blue (> 250), Red ($151 - 250$), Green ($76 - 150$), Magenta ($10 - 75$) ,Black ($1 - 10$)

3.3 The 1000 Genome Project and Population Stratification

3.3.1 Population Stratification and PCA

As introduced in Chapter 1, population stratification can have major effects on the detection of Linkage Disequilibrium, especially over long distances. Principle

Component Analysis, as described in the Appendix, was used to analyze the bi-allelic SNP genotype data from the 1000 Genomes Project for the presence of allele frequency differences between sub-populations. As suggested by Patterson et al., the rows the of genotype matrix being analyzed were indexed by individuals instead of by population labels, as is often done in PCA. One reason for such indexing is that population labels are inherently socially constructed, while individuals are not. In addition, indexing by individuals allows for the examination of within population variation as well as outliers. Further analysis can access the correlation of population labels to the variance structure determined by the eigenvector analysis [60].

In the study of the 1000 Genome Project data, PCA was implemented on the SNPs from chromosome 2 and chromosome 15, independently. Figure 3.3 is a plot of the coordinates of the top two eigenvectors for the genotype data from chromosome 2. Population separation is clear; however, the African subpopulation in black and the American subpopulation in red seem to appear in clines, as opposed to tight clusters around a central point [60].

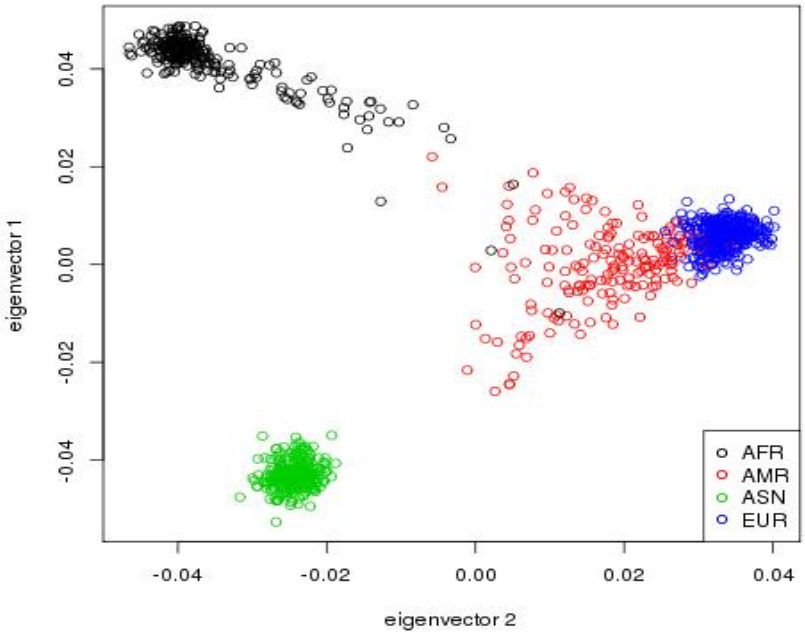


Figure 3.3: *Chromosome 2: Eigenvector 1 vs Eigenvector 2*

Figure 3.4 and Figure A.1 are the pairwise plots of the top four eigenvectors along, with their associated percentage of variance for each of the top eigenvectors. As you traverse down the list of eigenvectors, the separation between the population labels becoming increasingly less defined, as expected since less variance is explained by these eigenvectors.

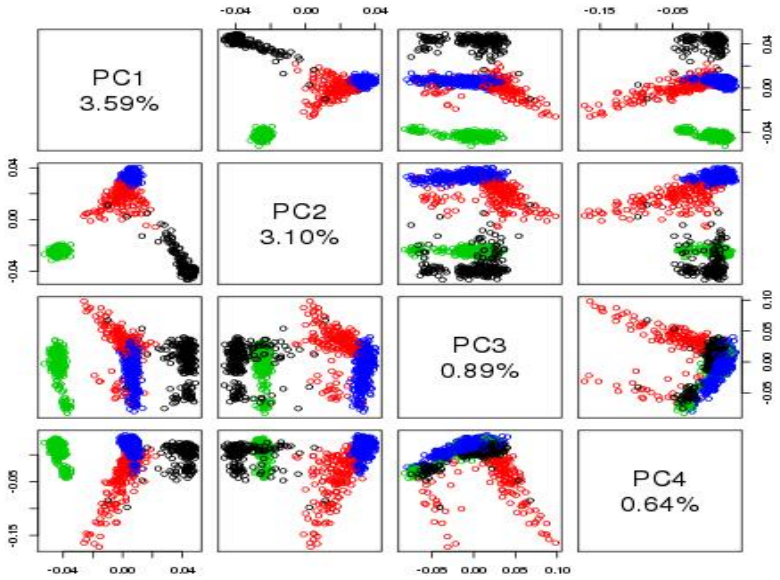


Figure 3.4: *Chromosome 2 Pairwise Plots of Top 4 Eigenvectors.* The pairwise plot of eigenvector coordinates is plotted for each pair of the top four eigenvectors. The points are color coded by their pre-specified population labels (AFR=black, AMR=red, ASN=green, EUR=blue). The diagonal boxes contain the percentage of variation accounted for by each eigenvector.

The top eigenvectors account for approximately one fifth of the sequence variation for the two chromosomes tested. ANOVA was used to test whether or not the subpopulation means for the top eigenvectors are significantly different and to access how much of the variation accounted for by the top eigenvectors can be attributed to the four major ancestral groups. The results from the ANOVA analysis are in Table 3.4 and Table A.1. For both chromosomes under investigation, the population means for the top 3 eigenvectors are statistically significantly different (p-value $< 10^{16}$). In addition, for the top two eigenvectors, for both Chr 2 and Chr 15, the proportion of variance explained by the 4 major subpopulation codes (R^2) is between .95 and .97. The results reflect that allele frequency differences between the populations account for a significant proportion of genomic variation and the need to analyze LD for the 4 four major ancestry groups independently of one another.

Number	Eigenvalue	F Statistic	p-Value
1	94.77	10404	$< 10^{-16}$
2	76.53	10426	$< 10^{-16}$
3	9.18	563.31	$< 10^{-16}$

Table 3.4: *1000 Genomes Chromosome 2 ANOVA.* Statistical analysis tests for significant differences in the means of the eigenvector coordinates between population groups determined by their pre-defined population labels.The top three eigenvectors were significantly significant, indicating correlation between the eigenvectors and population label.

3.3.2 Parameter Estimation by Subpopulation

For each of the four sub-populations, the parameter training procedure above was implemented on the same randomly sampled data sets on the individuals separated by the four ancestry population codes independently. Table 3.5 contains the averaged estimated model parameters for each of four major sub-populations:

Population	EP:U	EP:P	TP:N	TP:P
Afr	8.1361	2.6382 0.1396 0.0331	0.7469	0.3477
	1.4502	0.1304 0.9680 0.0576	0.1437	0.1967
	0.1439	0.0300 0.0517 0.2427	0.1094	0.2631
				0.1924
Eur	6.5039	4.3013 0.1202 0.0228	0.7304	0.5027
	1.3930	0.1195 2.1694 0.0549	0.1731	0.0740
	0.1848	0.0255 0.0535 0.4683	0.0965	0.3247
				0.0986
Asn	4.9613	4.0424 0.1212 0.0296	0.6446	0.4228
	1.3988	0.1158 2.1844 0.0512	0.2322	0.3008
	0.2379	0.0286 0.0493 0.4541	0.1233	0.0939
				0.1825
Amr	4.8858	8.2787 0.2386 0.0271	0.7485	0.4676
	1.2514	0.2410 4.0427 0.0758	0.2012	0.0898
	0.1814	0.0269 0.0844 0.7706	0.0503	0.3251
				0.1175

Table 3.5: *Averaged Training Parameters by Population* The averaged trained parameters over the four training repeats appear in the table. Starting from the third column, the table contains the emission parameters for state U, the emission parameters for the state P, the transition parameters for $N \to L|UL|U$ respectively and the transition parameters for $P \to PN|UL|U|P$, respectively

For the four populations, the Dirichlet parameters, specifically the means of the Dirichlet distributions, are relatively similar with no major statistical differences found. This result is consistent with previous studies which have suggested that allele frequencies are highly correlated among populations [61]. The results do show a slightly smaller variance in the distribution of unpaired SNPs as well as a slightly lower average minor allele frequency for people of African ancestry in comparison to the other three ancestry groups. In addition, for the African subpopulation, the variance of distribution of the paired SNPs is slightly higher in comparison to the other ancestry groups. Previous research has shown that individuals of African ancestry have greater genetic diversity and as a result a greater percentage of SNPs with major allele frequency greater than 0.95 when compared to the other ancestry groups [61]. Thus, the results from the *Inside-Outside Algorithm* for the SCFG are consistent with these previous findings. Finally, greater variance in distribution of the paired genotypes in the African population aligns with the fact that the power to accurately detect association is low for loci with small minimum allele frequencies [61].

It is also important to note that both modes of emissions of the genotypes are modeled by a hierarchical Multinomial with a Dirichlet prior. In previous research, attempts have been made to derive mathematical formulations for the spectra of expected allele frequency in genome wide SNPs data. Model fitting revealed population specific differences in demographic history which best describe the observed frequency spectra [62]. It was concluded that the African-American spectra shows history of moderate and uninterrupted population expansion, while the European and Asian spectra show evidence of a past reduction in population size, followed by a recent recovery in population size [62]. The former can successfully be modeled by a single Multinomial- Dirichlet distribution and the learned parameters are

consistent with the findings from this previous research. However, the reduction and subsequent grown of the European and Asian populations, otherwise known as a bottleneck, might be better modeled by a mixture of two Multinomial- Dirichlet distribtions, one modeling the allele frequencies during the expansion and the other modeling the allele frequencies during the collapse of the population size. In future research, it may be of interest to use alternative prior models to better reflect population history.

3.3.3 Association Inference for Multiple Genomic Regions by Subpopulation

The three genome segments of interest from the previous section were also analyzed for the separate ancestry populations. For each population, the joint probability of each genomic regions was calculated, given their respective trained model parameters. The centroid was determined based on back-samples derived from this distribution for each segment and for each subpopulation. The subsequent data analysis focuses on the African and European populations, as previous research has compared LD in these two populations. Table 3.6 contains the the number of pairs in the centroid for each of the genomic segments for the African and European populations, as well as the maximum length of all pairs in the centroids and the number of pairs with distance greater than 10kb. In addition, the eighth column of the table contains the number of pairs common to both populations and the final column is the number of common pairs that are a genomic distance greater than 10kb.

Firstly, the results indicate the existence of association between SNP loci extending beyond 5kb in the African population, with approximately 10 percent of the high

frequency pairs extending beyond 10kb. In addition, the data indicates that close to one third of the SNPs each population's centroid are common to both populations, leaving two thirds of the high frequency pairs differing between the populations, indicating the existence of both European and African specific associations. These two major results tend to conflict with some previous studies, such as that of Reich et al. in which it was concluded that for the African population they studied for a given region LD extended less than 5kb and that there was very little Nigerian specific LD. These conflicting results can be attributed to methodological differences between previous studies and the SCFG method. In the study of Reich et al, the length to which LD was measured was determined by the decay of LD and its half-life between a core SNP and SNPs in 2kb regions extending to 160kb, thus being unable to capture associations between distant non-contiguous regions and the variation in regions in which stronger LD exists at longer distances in regions where LD is weak over shorts distances [15]. In addition, LD was only measured from a single common core SNP (MAF $>$ .35), which is significantly higher than the MAF cutoff used in the SCFG model, limiting long range LD in the Reich study [15].

Comparison between the two centroids of the European and African population also shows a significant drop in the fraction of common pairs with LD extending to distance greater than 10kb (from 10 percent down to 1-5 percent). Positively, this result is consistent with previous findings that common linkage disequilibrium between the two populations is usually of short distance (i.e. local linkage due to lack of recombination) and it is indicative of common ancestry of about 100,000 years ago [15].

Chromosome	Afr	MD	$> 10kb$	Eur	MD	$> 10kb$	Af/Eur	$> 10kb$
2: 102319730	578	118k	80	533	120k	83	173	4
4: 104213866	746	247k	83	659	245k	58	248	1
18: 48742255	1177	527k	115	1123	192k	125	424	6

Table 3.6: *1000 Genomes Centroid Analysis by Population* This table contains the start position of the genomic segments, as well as the number of paired associations appearing in the centroid, the maximal distance between paired SNPs and the number of associated pairs of genomic distance greater than 10k for each population. It also contains the number of overlapping pairs from the centroids of each population.

Below are two pictorial representations of the associated pairs predicted in the centroids the their respective genomic regions. The segment of SNPs between the 250th and 500th sequenced SNP from Chr2:102319730 is depicted in Figure 3.5. The top image is the common ellipse diagram with the height of the ellipse in proportion to the number of SNPs in between the endpoints of the pair and the bottom is a scatter plot, where the points closest to the diagonal are indicative of recombination based local linkage and the points further from the diagonal indicate longer range linkage disequilibrium. Similar plots for the entire genomic segment Chr2: 10231973 and genomic segment Chr4: 104213866, both in its entirety and for the first 500 SNPs of the region appear in the Appendix.

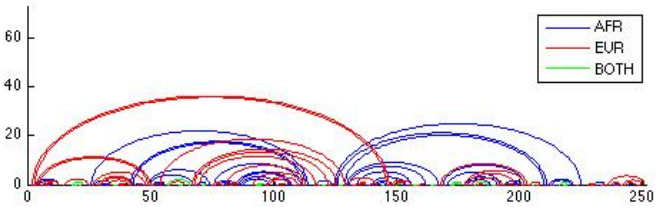
These plots visually confirm that the commonly shared LD between the African and European populations is almost always local and short range. In addition, it can also be seen that while some of the LD events predicted in the African population fall between longer range LD of the European sample, our results indicate that this is not always the case, with associated pairs predicted in the centroid for the African sample often encompassing regions pairs of shorter LD predicted in the centroid of the European sample. In addition, the data also shows that pairs predicted in their respective centroids are often in conflict between the two populations. This can be seen clearly in the Figure 3.5 and Figure A.6 as there are red arcs that cross blue arcs and there are red and blue arch which share only one endpoint in common.

Crossing arcs may be a reflection of historical and epistatic differences in the two subpopulations, but it may also be reflective of the constraint imposed by the SCFG model allowing only nested pairs. Non nested pairs cannot appear in the same structure, thus if a pair appears in the centroid of one subpopulation, the pair from the other subpopulation can not also appear in former populations centroid. Conflicting pairs of associated SNPs, which are each unique to the centroid of its subpopulation, where both left loci are within relatively short genomic distance and both the right loci are in relatively short genomic distance, may be indicative of common LD, with difference in loci resulting from the forces of local linkage and the constraints of the model. In addition, population specific associated pairs with only one loci in common may also be either biologically driven or reflective of the other constraint imposed by the SCFG model which restricts each based to be paired with at most one other base.

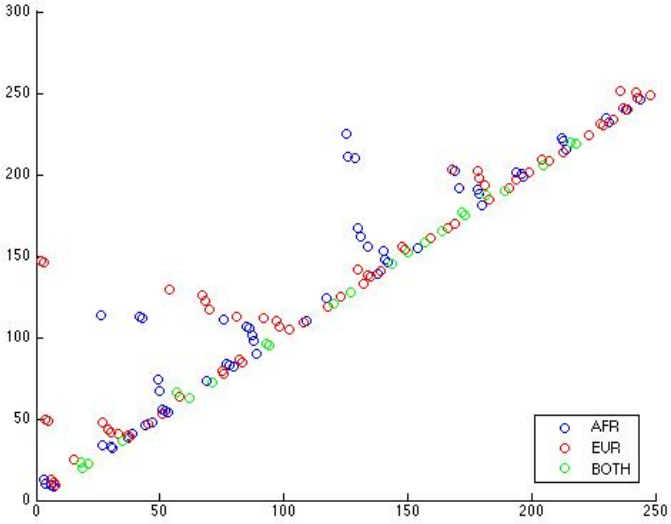
To investigate the effects of the model structure and constraints, the unique pairs in the centroid can be compared to the representative sample of the other subpopulation. For the genomic segment of Chromosome 2, of the 395 unique pairs in centroid of the African population, 49 appear in more than 10 percent of the representative samples, with 39 appearing in more than 25 percent of the representative samples of the European population. Also, of the 350 unique pairs in the centroid of the European population, 28 appear in more than 10 percent of the representative samples, with 15 appearing in more than 25 percent of the representative samples of the African population. Thus, for these pairs there is evidence that the SNPs are associated in both populations, with the differences in the populations being a marker of degree of association. It is important to note that all of these such pairs are local.

For predicted long range associations that are in conflict, their probabilities in

the alternate subpopulation are low to nonexistent and thus these differences can not be attributed to the SCFG model constraints, but rather reflect historical and epistatic differences between ancestry groups. These differences, whether historically or epistasis driven, suggest that differing haplotype blocks patters exist for each population and incorporation of the LD landscape from independent analysis in disease association studies may lead to increased power in accounting for phenotypic variations.



(a)



(b)

Figure 3.5: *Chr2: 102319730 Afr vs Eur (bases 250-500).* (a) The common ellipse diagram with the height of the ellipse in proportion to the number of SNPs in between the endpoints of the pair (b)scatter plot, where the points closest to the diagonal are indicative of recombination based local linkage and the points further from the diagonal indicate longer range linkage disequilibrium.

Side by side comparison between the centroids from the African population, the European population and the total ethnically diverse population, demonstrates that

the predicted associations differ greatly both between the groups and in comparison to analysis of the population as whole. When comparing centroids of the the genomic region of Chromosome 4, it appears that for some of the longer range associations that appear in the centroid of the total population, there are two shorter conflicting associated pairs in each of the two subpopulations. The region spanning from approximately the 150th loci to right after the 350th loci is an example of this, as well as the region from approximately the 500th loci to the 1000th loci. This is a great example of where population stratification has an effect on the prediction of LD, with some examples of long range LD found in the genetically diverse population, that are absent in the individual populations.

The results from PCA motivated the independent analysis of association in the subpopulations and the results of the statistical inference verify the importance of identification of population stratification prior to linkage and disease association analysis. The presence of allele frequency differences may produce spurious LD, and in these situations independent analysis may provide more power in accounting for phenotypic variation. In addition, when long range LD is detected in a subpopulations, it may be important to further investigate of the smaller ethnic groups to distinguish between the various causes/effects of long range LD.

3.4 An Application: The study of linkage disequilibrium in Preterm vs Full Term Pregnancies

3.4.1 Background and Motivation

Preterm birth, which is defined as birth prior to 37 weeks of gestation, is a common condition with incidence in the United States of at least 12% and is known to be on the rise [63]. Babies born prematurely may die neonatally and many others may be subject to an array of long-term side effects, thus creating "clinical, economic and psychological burdens" [64]. Women who have had a preterm pregnancy are at increased risk of having a subsequent early delivery, suggesting that preterm birth involves both genetic and environmental factors [65]. Identification of genetic susceptibility variants is of great interest in order to understand the pathogenesis of preterm birth. Current clinical tests and interventions demonstrate that single genes and pathways are inadequate to explain most preterm births [64]. With its increased prevalence and interest in investigating the genetic pathology, the results from a genome-wide study of preterm birth have been released from the Gene Environment Association Studies initiative (GENEVA). The data set includes the phenotype and genotype data from 4000 Danish women and children, with 1000 preterm mother/child pairs and 1000 full-term mother/child pairs. Genome-wide SNP genotyping was performed on the subjects (560,768 SNPs) and the data from this study is available in the Database of Genotypes and Phenotypes (dbGaP) [66]. Previous research has demonstrated that genetic risk of preterm birth is highly dependent on the maternal genotype only [67]; therefore, studies, including that of Uzun et al., have concentrate their analysis on the maternal genotypes only.

In their 2013 Genomic's paper, Uzun et al. studied preterm birth by identifying a set of genes for analysis that were validated by apriori biological information. Through data mining, natural language processing and pathway imputation, 617 genes were identified as having prior biological evidence for association to preterm birth. Single SNP association analysis was conducted on the 9077 tag SNPs that were found within the genomic region of these genes and within the regions 5kb upstream and downstream of each gene. The same association analysis was executed using all genotyped SNPs. It is also important to point out that preterm birth may be a quantitative trait, i.e. there may be varying susceptibility factors for different gestational categories [63]. Therefore, analysis was done dividing the women into three gestational age categories: less than 30 weeks, less than 34 weeks, and less than 37 weeks and comparing each category to the controls: greater than or equal to 38 weeks of gestation. There were no significant associations detected in the curated SNP list after correction for multiple comparison testing and there was only one SNP that achieved significance in the genome wide study, although it did not map back to a known gene [64]. These results are not surprising, given the success, or lack there of, of previous genome-wide single SNP based association studies. As an alternate approach to association analysis, Gene Set Enrichment Analysis was carried out on the SNPs from the curated gene set as well as genome-wide and looked for genes that were connected highly connected and for pathways that had the most genes contributing to their significance.

To investigate more complex associations, our SCFG was used to model both the cases and controls independently to make inference on jointly occurring pairs of pairs of SNPs in the curated SNP set. With the prior assumptions that linkage disequilibrium is non existent across chromosomes, the model was implemented on each chromosome independently.

3.4.2 Results and Discussion

The set of 9077 genotypes SNPs, from the curated gene set, were separated by chromosome. As motivated by the work of Uzun et al, the sampled genotypes of the approximately 2000 mothers were broken into categories determined by their completed number of gestational weeks. As mentioned above, preterm birth may in fact be a quantitative trait and to capitalize on the variance between gestational ages, individuals with gestational age of equal to or greater than 38 weeks were considered the control population and the individuals with gestational age of less than 34 weeks were consider the case population. *The Inside-Outside Algorithm* was implemented on the 813 SNPs from Chromosome 2, the 877 SNPs of Chromosome 6, and the 228 SNPs of Chromosome 22 for cases and controls independently. Representative samples were pulled from the joint probability obtained from analysis on each of the chromosomes and centroids were determined for each of the two populations. The number of pairs predicted in each of the centroids, as well as the number of predicted pairs in common between the centroids of the case population and the control population appears in Table 3.7.

Chromosome	Number of SNPs	Cases Only	Controls Only	Number of Common Pairs
22	228	10	19	74
6	877	82	104	254
2	813	39	58	258

Table 3.7: *Centroids of Preterm Birth Curated SNP Set by Chromosome* The number of predicted pairs for the case and control populations and the number of overlapping predicted pairs for the centroids of the SNPs from Chromsome 22, 6, and 2.

The case and control groups were defined as to obtain the greatest variation in the phenotype of the two populations . Both the cases and control populations are of Danish ancestry, limiting the overall genomic variance when compared to subpopulations of highly divergent ancestry. The results, as presented in Table 3.7, support this notion as the percentage of associated pairs in common between the cases and controls varies between 70-80%, whereas when when comparing ancestry populations, the percentage of common predicted pairs was only 30%.

The centroids of chromosomes 22, 6, and 2 are represented in Figure 3.5, with the arcs categorized as common to cases/controls (blue), present only in cases (green), and present only in controls (magenta). A majority of the overlapping predicted pairs are of short range (blue), indicative of the common ancestry of the two populations. The predicted pairs, which are specific to either the cases or the controls, fall into different categories with different biological implications.

First, as seen in both Chromosome 2 and Chromosome 6, many of the predicted pairs that are unique to the centroids of the cases or controls share one loci while the second differs, but are in close proximity (within the same gene) to each other. For example, in Chromosome 6, the SNP at the 223rd loci is paired to the 152nd loci in the controls more than 50 percent of representative samples, while it is paired to the 153rd loci in the controls more than 50 percent of the time (with both of these pairs mapping back to the gene pair HLA-B / VEGFA). In addition, six SNPs in the region of the 142-147 loci are paired to SNPs in the 763-877 loci region, all corresponding to the gene pair HLA-B/ PACRG, in both cases an controls, with different combinations of the endpoints. Seven of the high frequency pairs in the centroid of the controls also appear in the representative sample for the cases, with two of these pairs occurring in approximately one fourth of the representative samples of the cases. This class of conflicting pairings is also seen in Chromosome 2 with

predicted pairs in the centroid between regions around 320th loci and the 700-800th loci range. Thus such predicted pairs, while unique to each centroid, are reflective local linkage acting with the constraints of the model and differences can be viewed as markers of the extent of association.

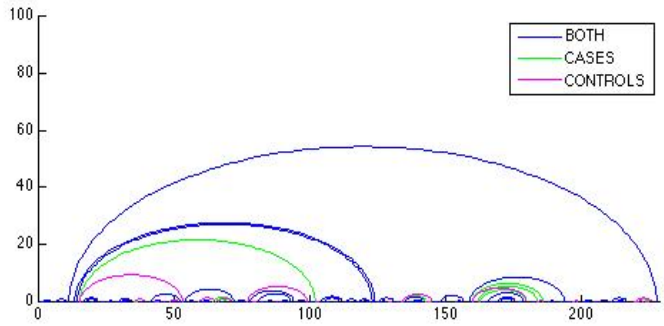
The effects of the model constraints can be accessed by comparing the unique centroid pairs to the lower frequency pairs in the opposing population's representative structure. For chromosome 6, 6 out of the 82 unique pairs in the centroid of the cases appear in the representative sample for the controls; however, they are all indicative of local linkage in a single gene. In addition, 15 out of the 104 unique pairs in the centroid for the controls appear in more than 10 percent of representative samples for the cases. Similarly 4 out of the 39 and 4 out of the 10 unique pairs in the centroid for the cases of Chromosome 2 and 22 respectively, occur in the representative sample of the controls. Finally, 13 out of the 58 and 10 of the 19 unique pairs in the centroid for the controls of Chromosome 2 and 22, respectively occur in the representative sample of the cases, with only 2 of the 19 pairs occurring in more than one fourth of the samples for Chromosome 22. The presence of these pairs can be explained by the model structure and constraints; however, all of the remaining unique pairs are indicative of potential epistatic interactions distinguishing the two phenotype classes.

It is also important to note that while the SCFG model does not permit a single loci to be paired to multiple other loci, the grammar models multiple pairs of pairs of SNPs. For example, we observe SNPs in the HLA-B gene associated to both SNPs in the VEGFA gene and the PACRG, reflecting more complex interactions extending beyond the realm of pairwise associations.

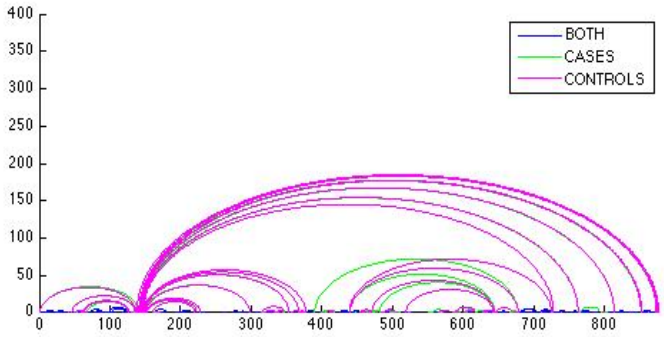
In the study of Chromosomes 2 , 6, and 22, the second class of cases/control

specific differences is such that the unique pairs from each group share one common loci, but the other loci localize to distinct regions of the genome. For example, in the centroid of Chromosome 22 of the cases, the SNP at the 15th loci is paired to the SNP at the 102nd loci. There is no predicated pair in the centroid of the control population such that the left loci close to SNP 15 and and the right close to SNP 102 simultaneously and the SNP pair (15,102) does not appear in the representative sample, thus indicative of potential epistatic interaction effecting the preterm phenotype. SNPs 15 and 102 are mapped to the GSTT2 gene and the MYH9 gene, with the former being a protein coding gene whose product catalyzes the conjugation of reduced glutathione and the latter is a protein coding gene whose production is non-muscle myosin. The presence of this association is novel when compared to the known gene ontology pathways; however, a pubmed search suggests that glutathione has some link to myosin function [68] A list of all the unique specific case/control gene pairs that are mapped from the high frequency jointly predicted SNP pairs for chromosome 2 and chromosome 22, appears in Table 3.8 and the gene pairs for chromosome 6 appear Table A.6.

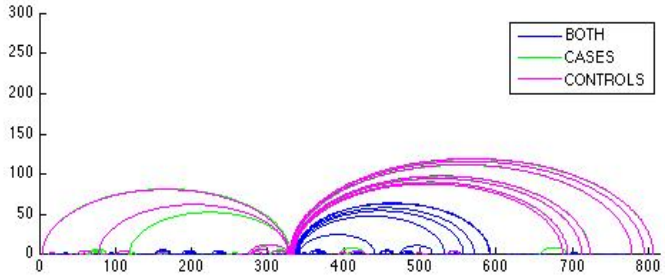
Further investigations into the biological relatedness of these two genes and other jointly occurring pairs of this nature, which are not predicted to be associated in the control population and vice versa, can provide insight into epistatic interactions resulting in increased risk for the disease, which when analyzed by independent association studies are not significantly associated to preterm birth or which may or may not be present in previously identified genomic pathways.



(a) SNPs of Chromosome 22



(b) SNPs of Chromosome 6



(c) SNPs of Chromosome 2

Figure 3.5: *Centroids of Curated Preterm Birth SNP Set by Chromosome* The blue ellipse are the predicted pairs in the centroids common to both the cases and controls, the green are the pairs specific to the cases, and the magenta are the pairs specific to the controls (a) The sequence of 228 pre curated SNPs from Chromosome 22 were analyzed, with 74 predicted pairs in common in the centroids of the cases and controls (b)The sequence of 877 pre curated SNPs from Chromosome 6 were analyzed, with 254 predicted pairs in common in the centroids of the cases and controls (c)The sequence of 813 pre curated SNPs from Chromosome 2 were analyzed, with 258 predicted pairs in common in the centroids of the cases and controls

Chromosome	Gene1	Gene2	Case/Control Only
22	GSTT2	MYH9	Cases
22	MLC1	PANX2	Cases
22	RBM9	MYH9	Controls
22	GSTT2	TIMP3	Controls
22	MLC1	PANX2	Controls
22	MLC1	MAPK11	Controls
2	CEBPZ	THADA	Cases
2	ALK	THADA	Cases
2	POMC	THADA	Cases
2	COL3A1	TMEFF2	Cases
2	THADA	TMEFF2	Cases
2	THADA	UGT1A1	Cases
2	THADA	RAMP1	Cases
2	ALK	THADA	Controls
2	POMC	THADA	Controls
2	THADA	TMEFF2	Controls
2	THADA	RAMP1	Controls

Table 3.8: *Inferred Epistatic Gene-Gene Associations.* Contains the Gene-Gene Pair for the unique paired SNPs in the Centroids of Cases or Controls. The only gene pairs that appear in the list are the pairs in with the paired SNPs are associated to different genes. All repeats pairs, mapping back to the same gene-gene pair are excluded.

Chapter Four

Future Directions

4.1 Future Directions: Extensions of the SCFG

Our SCFG model, in its current form, is designed to model simultaneous interactions between two SNP loci. While such analysis is more powerful than simple pair-wise LD estimates, association is biologically not limited to pairs of SNPs. To fully understand more complex interactions both locally (aka haplo-blocks /diplo-blocks) and distantly, it will be important to be able to model association between multiple loci. In addition, it is of great interest to be able to detect associations between genetic variants and disease phenotypes in order to profile disease manifestations, predict disease risk, and for potential treatment targets. Below, we will present possible extensions to the current SCFG model to address these two biological areas interest.

4.1.1 From Pairs to Triplets

It is possible to model long range interactions between multiple SNPs by adding in some additional states to the non-terminal and terminal state spaces to our current SCFG model. For example, we can model three way interactions by adding the following non-terminal states to the grammar $\boldsymbol{G}$:

- T: a string with the left most SNP associated to the rightmost SNP *and* exactly one other SNP in the string
- TS: a bifurcated string with the SNPs up to and including the bifurcation point being generated by state T and the SNPs beyond the bifurcation point generated by state S
- TN: a bifurcated string with the SNPs up to and including the bifurcation

point being generated by state T and the SNPs beyond the bifurcation point generated by state N

Just as non-terminal states P and U emit the summary statistics S_{ij} and S_i respectively, non-terminal state T emits a newly defined summary statistic S_{ikj}. S_{ikj} is the 3 dimensional matrix of genotype counts for the respective triplet of SNPs (i, k, j). In addition, note the changes (in bold font) to the production rules, $\boldsymbol{R}$, of the grammar:

$$
\begin{aligned}
S &\rightarrow N \mid E \\
N &\rightarrow L \mid UL \mid U \\
L &\rightarrow PS \mid \boldsymbol{TS} \\
PN &\rightarrow P/N \\
PS &\rightarrow P/S \\
\boldsymbol{TS} &\rightarrow \boldsymbol{T/S} \\
\boldsymbol{TN} &\rightarrow \boldsymbol{T/N} \\
UL &\rightarrow U/L \\
U &\rightarrow xU \mid xE \\
P &\rightarrow yPy \mid yUy \mid yPNy \mid yULy \mid \boldsymbol{yTy} \mid \boldsymbol{yTNy} \mid yEy \\
\boldsymbol{T} &\rightarrow \boldsymbol{zSzSz} \\
E &\rightarrow \epsilon
\end{aligned}
$$

With the aforementioned grammatical changes, for a given sequence of SNPs from i to j, one could compare both the probability of i and j being in disequilibrium

with each other versus the probability of these loci being jointly in disequilibrium with a third SNP somewhere in between them. It is important to note the the later probability is the sum of the probabilities over all possible positions of the third associated loci. In addition, the new terminal state T can itself be viewed as a bifurcating state because, like the other bifurcation states, it produces two substrings. However, in comparison with all other bifurcating states, the two substrings result from the emission of a three loci interaction.

A modified *Inside and Outside Algorithms* can be applied to this extended grammar and each algorithm can be computed in $O(n^3)$ time in both number of non-terminal states and length of query sequence. This grammar has the same order of complexity as as the original SCFG model because the new non-terminal state T produces at most two substrings, i.e.. transitions to two non-terminal states and thus upper bound on the computation time is $O(n^3)$ in the number of non-terminal states. While the order of the time complexity will not change, the drawback of these grammatical changes is the lack of a straightforward 2d representation and increase in the number of transition and emission parameters that must be estimated. There is always a tradeoff as to the amount of information that can be gained and the computationally complexity that come along with such gains. Implementing these grammatical changes and investigating this trade-off is of future interest.

4.1.2 The SCFG Model & GWAS

Genome Wide Association Studies were introduced into the scientific community over a decade ago with the potential to be a powerful tool in uncovering the genetic basis of disease [69]. GWAS can be viewed as an approach to query the genome with the goal of identifying novel genetic variants for complex traits [70]. There are various

methods used to analyze such data, and while studies have have discovered some important genetic variants in regards to disorders such as macular degeneration, inflammatory bowel disease and obesity, others have failed to detect variants that account for phenotypic variation.

The most commonly used methods for association studies include single marker SNP or haplotype analysis. The former method suffers from a lack of statistical power due to the presence of LD which is not accounted for during analysis or during the correction for multiple comparisons [70]. In an attempt to include local LD information, haplotype analysis methods have been developed, which are potentially more powerful than single SNP methods, but are also computationally more intensive [71]. Another downside to these methods is that haplotypes are not directly observed, but rather need to be inferred via a phasing procedure. This can result in biased estimators and a decrease in power [72, 10]. To address such issues, the sliding window approach was developed, in which the likelihood ratio statistic for a given window would account for phase uncertainty and sampling [73]. Such analysis can result in a loss of power due to very high degrees of freedom, the fact that single sized windows are not globally optimal and that windows ignore recombination hotspots [74]. Haplotype block analysis has replaced the sliding window approach, but as mentioned perviously in the discussion of LD mapping, block definition is often arbitrary and LD may exist between blocks [75].

Recent methods attempt to improve association studies by incorporating LD information into the analysis. For example, Li et al. proposed to incorporate LD information into the identification of disease associated SNPs via a Markov random field model (HMRF). Using a weighted LD graph, LD is incorporated into the calculation of the posterior probability of a SNP being associated with the disease [70]. This method, although demonstrated to be more powerful than the previous single

SNP analysis, requires a previously computed LD graph and the authors only used the posterior probability to make inference on individual SNP associations. In the second method, Zhang et al. both incorporate LD and allow for epistatic interactions between SNPs and the disease phenotype. They simultaneously infer blocks of linkage disequilibrium on the genotype data while searching for SNPs within and between blocks that are associated to disease both independently and jointly via a Monte Carlo Markov Chain [10]. A Metropolis Hasting sampling method is used to iteratively assign SNPs to nonempty blocks and to assign association categorization to the SNPs (not associated to disease, independent and association, epistatic and associated). As mentioned in a previous section, this algorithm uses MCMC to approximate the posterior distribution conditional on the data due to the fact that the normalizing constant is too complex to calculate [10].

It may be possible to simultaneously make inference on LD and both single disease associations as well as paired associations without needing a sampling procedure to approximate the distribution via an extension of our Stochastic Context Free Grammar Model. The SCFG model allows for the calculation of the normalizing constant in $O(n^3)$ time and for direct sampling from the joint distribution.

Consider the following changes to the data: the data matrix $S_{<1,L>}$ will be an $M \times (L+1)$ matrix where M is the total number of individuals in the sample and L is the number of SNPs genotyped. The first L columns are the genotypes for each individual for each SNP and the L+1 column takes values $\{0, 1\}$ indicating the absence or presence of the phenotype of interest. The data matrix for any subsequence of SNPs, $S_{<i,j>}$, is also augmented with this phenotypic column. Below is the extension of the example from Chapter 2 to include phenotypic information:

ind/snp	1	2	3	4	5	6	7	phenotype
n_1	0	0	2	1	1	0	1	0
n_2	1	0	0	0	1	1	1	0
n_3	0	1	0	2	0	0	0	1
n_4	1	1	2	1	0	2	0	0
n_5	2	0	1	0	2	1	2	1

In addition to the two summary statistics/ terminal states S_i and S_{ij} that are already defined, this augmented model will require two additional statistics: S_i^A and S_{ij}^A. The former is a 3×2 matrix of genotype/phenotype counts for SNP at position i and the latter is a $3 \times 3 \times 2$ matrix of paired genotypes/phenotype counts for the pair of SNPs at position i and j. For example: $S_1^A = \begin{matrix} 1 & 1 \\ 2 & 0 \\ 0 & 1 \end{matrix}$, where $S_1^A(1,1)$ is the number of sampled individuals without the phenotype of interest that have 0 minor alleles for SNP loci 1 and $S_1^A(1,2)$ is the number of sampled individuals with the phenotype that have 0 minor alleles, and so on and so forth.

In addition, note the following necessary changes (in bold) to the production rules, $\boldsymbol{R}$:

$$
\begin{aligned}
S &\rightarrow N \mid E \\
N &\rightarrow L \mid \boldsymbol{U^*L} \mid \boldsymbol{U^*} \\
L &\rightarrow PS \mid \boldsymbol{P^AS} \\
PN &\rightarrow P/N \\
PS &\rightarrow P/S \\
\boldsymbol{P^AS} &\rightarrow \boldsymbol{P^A/S} \\
\boldsymbol{P^AN} &\rightarrow \boldsymbol{P^A/N} \\
\boldsymbol{U^*L} &\rightarrow \boldsymbol{U^*/L} \\
\boldsymbol{U^*} &\rightarrow \boldsymbol{U|U^A} \\
U &\rightarrow \boldsymbol{xU^*} \mid xE \\
\boldsymbol{U^A} &\rightarrow \boldsymbol{xU^*} \mid \boldsymbol{xE} \\
P &\rightarrow xPy \mid xPNy \mid \boldsymbol{xP^Ay} \mid \boldsymbol{xU^*y} \mid \boldsymbol{xP^ANy} \mid \boldsymbol{xU^*Ly} \mid xEy \\
\boldsymbol{P^A} &\rightarrow \boldsymbol{xPy} \mid \boldsymbol{xPNy} \mid \boldsymbol{xP^Ay} \mid \boldsymbol{xU^*y} \mid \boldsymbol{xP^ANy} \mid \boldsymbol{xU^*Ly} \mid \boldsymbol{xEy} \\
E &\rightarrow \epsilon
\end{aligned}
$$

where the non-terminal U^* is any string of unpaired elements, U is a string of unpaired SNPs with the left most being independent of phenotype, and U^A is a string of unpaired SNPs with the left most being associated to given phenotype. Similarly, P is a string of SNPs with the left and right most associated to each other, but independent of phenotype and P^A is a string of SNPs with the left and right most associated to each other and to the disease phenotype. Additionally, just as before, states U and P emit S_i and S_{ij} respectively and states U^A and P^A emit S_i^A and S_{ij}^A. Assumed that for SNP i, the genotype frequencies for individuals with disease

phenotype (cases) have the following probabilities (p_0, p_1, p_2) and for the controls the frequencies have the probabilities $(\theta_0, \theta_1, \theta_2)$ and both probability vectors are distributed from a prior dirichlet distribution with parameters $(\alpha_0, \alpha_1, \alpha_2)$. If SNP i is not associated to the disease phenotype, the cases should have the same frequency as the controls and thus:

$$\begin{aligned} Prob(S_i^A|U) &= \int Prob(S_i^A(:,1) + S_i^A(:,2)|\vec{p_i})Prob(\vec{p_i})d(\vec{p_i}) \\ &= \frac{\Gamma(\sum_{k=0:2}\alpha_k)\prod_{k=0:2}\Gamma(\alpha_k + n_{k,1}^i + n_{k,2}^i)}{\prod_{k=0:2}\Gamma(\alpha_k)\Gamma(\sum_{k=0:2}\alpha_k + \sum_{k=0:1} n_{k,1}^i + n_{k,2}^i)} \end{aligned}$$

where $n_{k,1}^i$ is the number of controls with genotype k and $n_{k,2}^i$ is then number of cases with genotype k. The above emission probability is the same as the emission probability from non-terminal state U in the original grammar, $Prob(S_i^A|U) = Prob(S_i|U)$. The same is true for the pair of associated SNPs, i and j. That is, $Prob(S_{ij}^A|P) = Prob(S_{ij}|P)$.

If SNP i is associated to the disease phenotype, the cases will not have the same probability as the controls:

$$\begin{aligned} Prob(S_i^A|U^A) &= \int Prob(S_i^A(:,1)|\vec{\theta_i})Prob(\vec{\theta_i})d(\vec{\theta_i}) \times \int Prob(S_i^A(:,2)|\vec{p_i})Prob(\vec{p_i})d(\vec{p_i}) \\ &= \frac{\Gamma(\sum_{k=0:2}\alpha_k)\prod_{k=0:2}\Gamma(\alpha_k + n_{k,1}^i)}{\prod_{k=0:2}\Gamma(\alpha_k)\Gamma(\sum_{k=0:2}\alpha_k + \sum_{k=0:1} n_{k,1}^i)} \times \frac{\Gamma(\sum_{k=0:2}\alpha_k)\prod_{k=0:2}\Gamma(\alpha_k + n_{k,2}^i)}{\prod_{k=0:2}\Gamma(\alpha_k)\Gamma(\sum_{k=0:2}\alpha_k + \sum_{k=0:1} n_{k,2}^i)} \end{aligned}$$

A similar expression exists for $Prob(S_{ij}^A|P^A)$, where probability of the genotypes will be a nine category multinomial and the prior distribution will be a nine category dirichlet. A modified version of *The Inside-Outside Algorithm* can be used on this

augmented grammar to compute the probability of the genome sequence given the model and back sampling from the ensemble of possible parses will provide a means to make inference about jointly occurring LD between SNPs, multiple single SNPs associated to disease and pairs of pairs of SNPs associated disease. This model has the potential to detect various modes of disease association/risk and the potential to gain predictive power from the joint probability of association events in comparison to previous studies that take each association independently of one another.

Appendix A

Supplemental Methods and Figures

A.1 Principal Component Analysis for Genomic Data

Following the methodology of Price et al. [31], Principal Component Analysis was implemented on the data from the 1000 Genome Project as follows:

Let M represent the number of individuals sampled and let L represent the number of SNPs sequenced for each individual. We will represent the genotype of SNP j for individual m as $s_{m,j} = \{0, 1, 2\}$. The integer value of the genotype represents the number of alternative alleles at that locus. The first step in Principal Component Analysis of the genotype matrix $G = S_{<1,L>}$ is to center the genotype matrix via subtraction of the column means $(u_1, u_2 \dots u_l)$. The centered matrix, C, is then normalized to assure that each data column has the same variance. Next, singular value decomposition is computed on C or the normalized matrix C_{norm}. For SVD, the $(M \times M)$ covariance matrix is $X = \frac{1}{l}CC'$ and the eigenvalue decomposition is computed for X. The eigenvalue decomposition will produce matrix A where the columns correspond to the ordered eigenvectors. These columns are known as the loading vectors.

The projection of the original data $G = S_{<1,L>}$ onto the space of eigenvectors is known as the score and is computed as follows:

$$Score = A^T \times C$$

The Score matrix has M rows and L columns. The score, s_{ij} for a SNP i at the

jth component is thus the dot product between the M-dimensional centered genotype vector ($\vec{c_i}$) and the transpose of the$M$-dimensional eigenvector ($\vec{a_j}$). The centered data can be reconstructed by:

$$\begin{aligned} DataCentered &= FeatureVectors \times Scores \\ &= A \times (A^T \times C) \end{aligned}$$

Thus for individual n at SNP j the reconstructed original genotype data is :

$$\begin{aligned} c_{m,j} &= \sum_{p=1:M} (a_{mp} \sum_{n=1:M} (a_{np} \times c_{nj})) \\ s_{mj} - u_j &= \sum_{p=1:M} (a_{mp} \sum_{n1:M} (a_{np} \times c_{nj})) \end{aligned}$$

Thus we can get the original data (uncentered) by taking the reconstructed data and adding back the column mean:

$$\begin{aligned} OrignalData &= (FeatureVector \times Scores) + OriginalMean \\ s_{mj} &= \sum_p (a_{mp} \sum_n (a_{np} \times c_{nj})) + u_j \end{aligned}$$

If we were to only select the top k principal components based on the amount of variance the k components explain, then in the reconstruction of the data above the sum over the principal components p will range from $p = 1 : k$. This reconstructed data would be an approximation to the original data. The difference between the original data and the k principal component- reconstructed data is known as the

residual. In his paper, Price refers to the difference between the original data and the k-component reconstructed data as "adjusted data" and uses it as the basis for association analysis.

$$s_{nj,adjusted} = s_{nj} - u_j - \sum_{p=1}^{k}(a_{np} \sum_{m}(a_{mp} \times c_{mj}))$$

Association analysis can then be computed on this adjusted genotype data with adjustments to the emission distributions. Using the adjusted genotype data would call for different emission parameters as the space is not longer a 3 category multinomial.

A.2 Principal Component Analysis for Chromosome 15

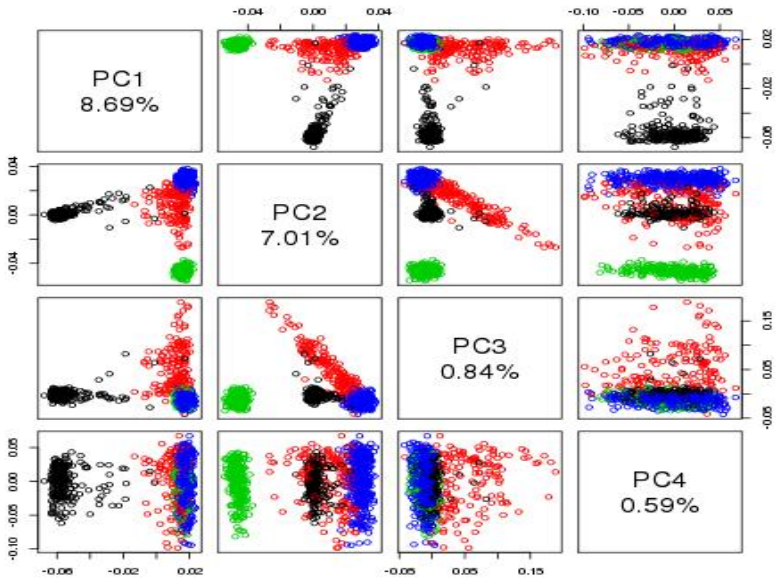


Figure A.1: *Chromosome 15 Pairwise Plots of Top 4 Eigenvectors.*The pairwise plot of eigenvector coordinates is plotted for each pair of the top four eigenvectors. The points are color coded by their pre-specified population labels (AFR=black, AMR=red, ASN=green, EUR=blue). The diagonal boxes contain the percentage of variation accounted for by each eigenvector

Number	Eigenvalue	F Statistic	p-Value
1	95.12	7824	$< 10^{-16}$
2	71.10	8259	$< 10^{-16}$
3	8.18	424.62	$< 10^{-16}$

Table A.1: *1000 Genomes Chromosome 15 ANOVA*. Statistical analysis was done to test for significant differences in the means of the eigenvector coordinates between population groups determined by their pre defined population labels. The top three eigenvectors were significantly significant, indicating correlation between the eigenvectors and population label.

A.3 1000 Genomes Centroid Predictions Cont'd

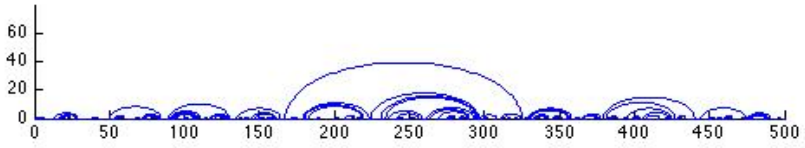


Figure A.2: *Chromosome 4: 104213866 (SNPs 1-500)*

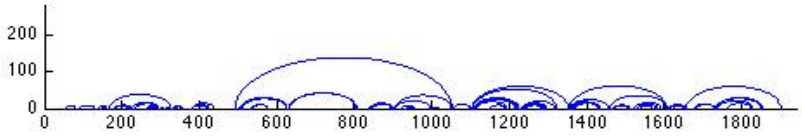


Figure A.3: *Chromosome 4: 104213866*

A.4 1000 Genomes Training Data By Population

Repeat	EP:U	EP:P	TP:N	TP:P
1	9.4088 1.5396 0.1667	3.0173 0.1597 0.0391 0.1282 1.0557 0.0786 0.0256 0.0565 0.3600	0.7013 0.1079 0.1907	0.2324 0.2176 0.3362 0.2138
2	5.6313 1.1373 0.1084	2.6890 0.1402 0.0307 0.1410 0.9863 0.0500 0.0316 0.0552 0.2206	0.7016 0.1705 0.1279	0.4556 0.2476 0.1436 0.1531
3	9.2596 1.6198 0.1566	2.6513 0.1314 0.0336 0.1362 0.8899 0.0474 0.0350 0.0483 0.1716	0.6820 0.2306 0.0874	0.4207 0.1234 0.2161 0.2398
4	8.2446 1.5041 0.1440	2.1951 0.1272 0.0291 0.1164 0.9400 0.0546 0.0278 0.0468 0.2184	0.9026 0.0658 0.0316	0.2822 0.1984 0.3565 0.1629

Table A.2: *1000 Genome Parameter Training : African*

Repeat	EP:U	EP:P	TP:N	TP:P
1	7.1710 1.3558 0.1527	2.3439 0.1018 0.0160 0.0970 1.2443 0.0440 0.0215 0.0404 0.3567	0.8686 0.1314 0	0.4668 0 0.2816 0.2516
2	5.4543 1.2081 0.1726	2.3468 0.1039 0.0225 0.1127 1.0472 0.0434 0.0316 0.0481 0.2466	0.7151 0.2849 0	0.1429 0.1429 0.5714 0.1428
3	5.1992 1.3113 0.2302	9.9282 0.1608 0.0354 0.1689 4.9995 0.0776 0.0308 0.0816 0.9559	0.5882 0.1251 0.2867	0.5449 0.1532 0.3019 0
4	8.1912 1.6969 0.1838	2.5862 0.1143 0.0173 0.0995 1.3868 0.0547 0.0180 0.0441 0.3142	0.7498 0.1508 0.0994	0.8561 0 0.1439 0

Table A.3: *1000 Genome Training Parameters : European*

Repeat	EP:U	EP:P	TP:N	TP:P
1	4.2826	1.3581 0.0920 0.0307	0.9237	0.6352
	1.1946	0.0958 1.0000 0.0423	0	0.3648 0
	0.2020	0.0290 0.0475 0.2991	0.0763	0
2	5.9345	2.5721 0.1041 0.0393	0.4000	0.5000
	1.3949	0.0925 1.3937 0.0499	0.4134	0.5000 0
	0.2269	0.0312 0.0372 0.2857	0.1866	0
3	5.3552	2.2546 0.0800 0.0195	0.6597	0.3976
	1.3562	0.0870 1.3348 0.0381	0.1101	0.0688
	0.2025	0.0191 0.0385 0.3790	0.2302	0.2453
				0.2883
4	4.2728	9.9849 0.2086 0.0287	0.5949	0.1584
	1.6494	0.1880 5.0090 0.0746	0.4051	0.2696
	0.3203	0.0350 0.0742 0.8524	0.0000	0.1303
				0.4417

Table A.4: *1000 Genome Training Parameters: Asian*

Repeat	EP:U	EP:P	TP:N	TP:P
1	3.9293	9.9389 0.2621 0.0219	0.8084	0.4172 0
	1.2174	0.2724 5.0800 0.0814	0.1916 0	0.2646
	0.2061	0.0224 0.0943 1.0429		0.3181
2	5.4003	3.1350 0.1309 0.0290	0.8762	0.3336
	1.1824	0.1304 1.1496 0.0436	0.1238	0.1658
	0.1335	0.0256 0.0475 0.1998	0.0000	0.5006
				0
3	4.4218	10.0034 0.2618 0.0303	0.5085	0.4159
	1.1723	0.3000 4.9925 0.0944	0.3402	0.1936
	0.1931	0.0340 0.1083 0.9514	0.1513	0.3905
				0
4	5.7917	10.0376 0.2998 0.0272	0.8008	0.7037 0
	1.4335	0.2614 4.9487 0.0838	0.1494	0.1445
	0.1929	0.0258 0.0873 0.8885	0.0498	0.1518

Table A.5: *1000 Genome Training Parameters: American*

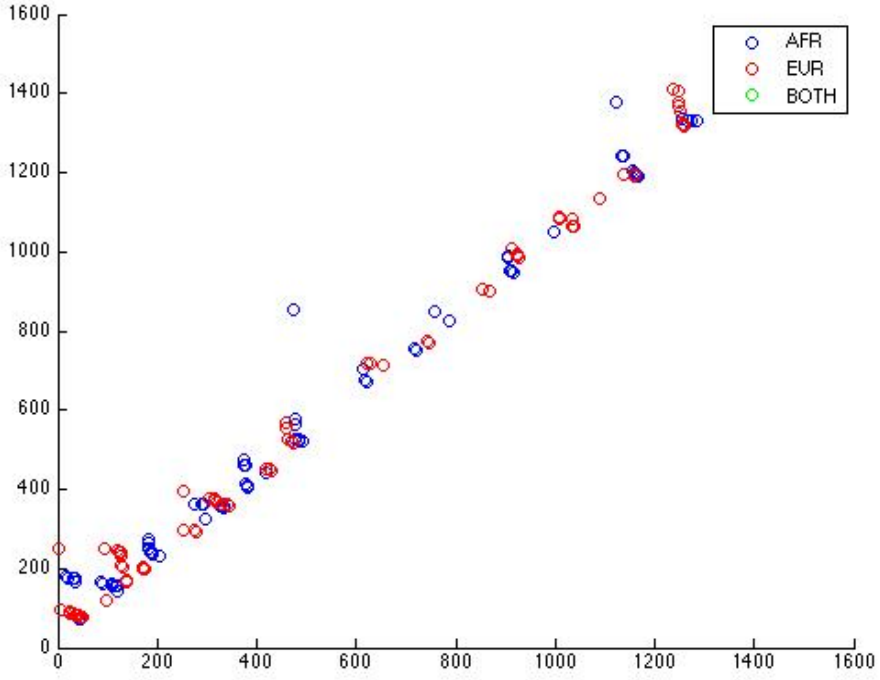


Figure A.4: *Chromosome 2: 102319730 African vs European Scatter Plot*

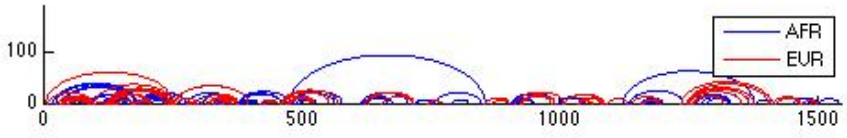


Figure A.5: *Chromosome 2: 102319730 African vs European*

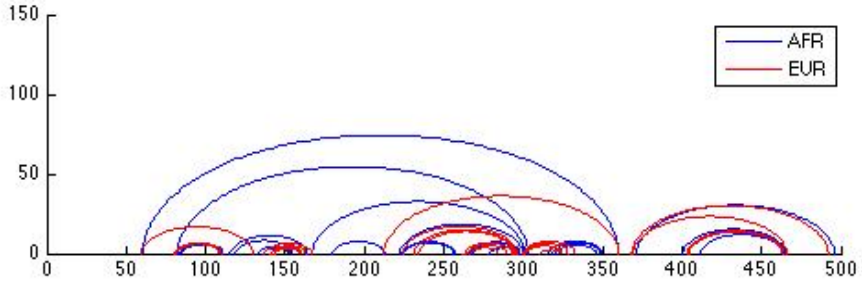


Figure A.6: *Chromosome 4: 104213866 African vs European (SNPs 1-500)*

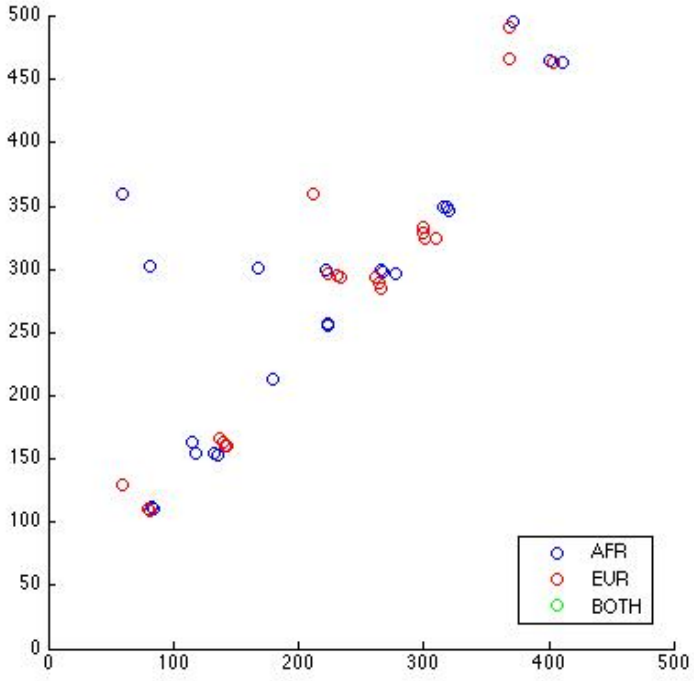


Figure A.7: *Chromosome 4: 104213866 African vs European Scatter Plot (SNPs 1-500)*

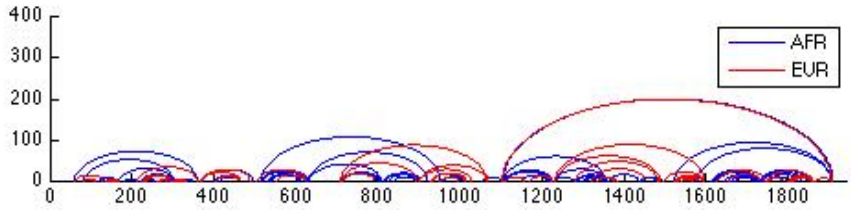


Figure A.8: *Chromosome 4: 104213866 African vs European*

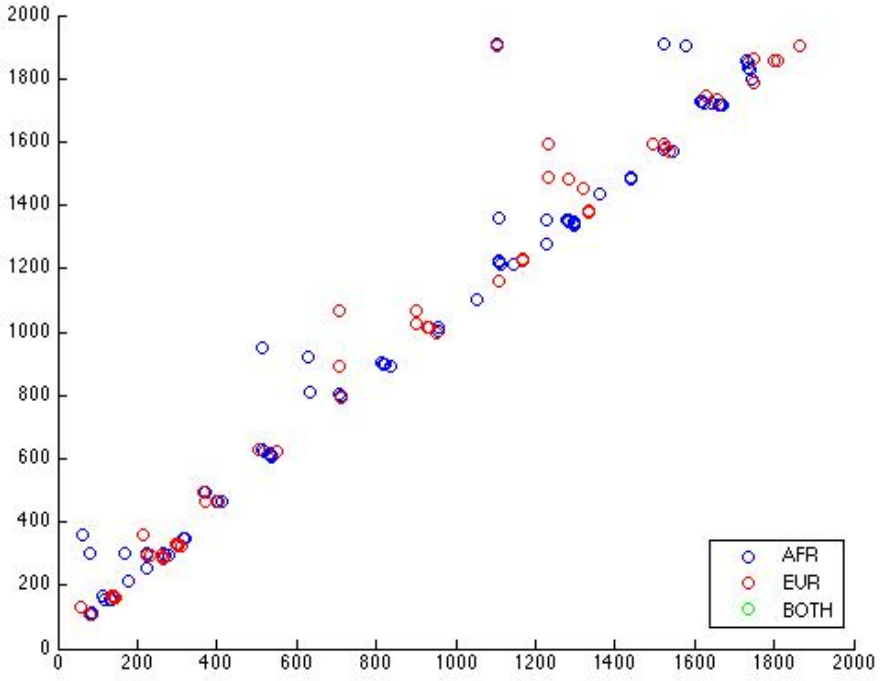


Figure A.9: *Chromosome 4: 104213866 African vs European Scatter Plot*

A.5 Preterm Birth Unique Gene-Gene Associations

Chromosome	Gene1	Gene2	Case/Control Only
6	F13A1	ID4	Cases
6	PRL	HLA-B	Cases
6	WRN1P1	HLA-B	Cases
6	HLA-B	PTCRA	Cases
6	HLA-B	VEGFA	Cases
6	HLA-B	NFKB1E	Cases
6	HLA-B	DDO	Cases
6	AKAP12	UST	Cases
6	HLA-B	HS3ST5	Cases
6	UST	ESR1	Cases
6	UST	PLG	Cases
6	UST	PACRG	Cases
6	F13A1	ID4	Controls
6	F13A1	HLA-B	Controls
6	WRN1P1	HLA-B	Controls
6	HSPA1A	PTCRA	Controls
6	HLA-B	VEGFA	Controls
6	HLA-B	NFKB1E	Controls
6	HLA-B	DDO	Controls
6	FYN	F13A1	Controls
6	HLA-B	HS3ST5	Controls
6	AKAP12	ESR1	Controls
6	UST	ESR1	Controls
6	UST	PLG	Controls
6	UST	PACRG	Controls
6	HLA-B	PACRG	Controls

Table A.6: *Inferred Epistatic Gene-Gene Associations: Chromosome 6.* Contains the Gene-Gene Pair for the unique paired SNPs in the Centroids of Cases or Controls. The only gene pairs that appear in the list are the pairs in with the paired SNPs are associated to different genes. All repeats pairs, mapping back to the same gene-gene pair are excluded.

Appendix B

Glossary of Notation

B.1 Glossary of Notation

- M: number of individual whose genomes were sampled
- L: number of SNP loci sequenced for each individual genome
- $s_{m,j}$: genotype at SNP locus j for individual m, number of alternative alleles at the locus
- $S_{<i,j>}$: matrix of genotype data for a subsequence of SNPs from the i^{th} locus to the j^{th} locus.
- S_i: summary statistic/terminal state, vector of genotype counts for SNP i
- S_{ij}: summary statistic/terminal state, matrix of genotype counts for pair of SNPs i and j
- n_{k^i}: number of times genotype k appears at SNP locus i
- $n_{k^i l^j}$: number of times genotype kl appears for SNP loci i and$j$
- $\boldsymbol{G}$: the Stochastic Context Free Grammar model
- $\boldsymbol{A}$: non-terminal state space of grammar
- $\boldsymbol{\Sigma}$: terminal state space of grammar
- $\boldsymbol{R}$: set of production rules for the grammar
- $\boldsymbol{\Theta}$: model parameters of the SCFG consisting for transtion and emission parameters
- $t(V, Y)$: transtion probability from non-terminal state V to non-terminal state Y

- $emit(x|V)$: probability of emitting terminal state x given x is rooted in non-terminal state V
- $\vec{p}$: vector of parameters of the multinomial distribution
- $\vec{\alpha}$, $\vec{\beta}$: vector of parameters of dirichlet distributions for states P and U respectively
- W_P : non-terminal states to which non-terminal state P can transtion to $\{PN, UL, U, P\}$
- W_N: non-terminal states to which non-terminal state N can transtion to $\{L, UL, U\}$
- Δ_v^L: number of terminal states emitted to left by non-terminal state v
- Δ_v^R: number of terminal states emitted to right by non-terminal state v
- N_i: total genotype counts for SNP i, should equal M
- $N_{<i,j>}$:total gentopype counts for SNP sequence i to j, should equal M
- $in(i, j, v)$: probability of subsquence of SNPs i to j and their genotype counts given subsequence is rooted in state v
- $out(i, j, v)$: probability of entire sequence of SNPs, exlcuding SNP i to j and their genotype counts which are rooted in state v
- $\boldsymbol{H}$: hessian of the log likelihood of the snps given the emission parameters
- $\boldsymbol{g}$: gradient of log likelihood of snps given the emission parameters
- G : $M \times L$ matrix of genotypes (indivuduals x snps)
- C: Centered matrix of genotypes

- A: Matrix of ordered eigenvectors
- s_{ij}: projection of original genotype data for snp i onto j^{th} eigenvector
- $g_{n,j,adjusted}$: k-component reconstructed genotype data for snp j for individual n

Bibliography

[1] Sachidanandam, R, Weissman, D, Schmidt, S, Kakol, J, Stein, L, Marth, G, Sherry, S, Mullikin, J, Mortimore, B, Willey, D, Hunt, S, Cole, C, Coggill, PC. nad Rice, C, Ning, Z, Rogers, J, Bentley, D, Kwok, P, Mardis, E, Yeh, R, Schultz, B, Cook, L, Davenport, R, Dante, M, Fulton, L, Hillier, L, Waterston, R, McPherson, J, Gilman, B, Schaffner, S, Van Etten, W, Reich, D, Higgins, J, Daly, M, Blumenstiel, B, Baldwin, J, Stange-Thomann, N, Zody, M, Linton, L, Lander, E, & Altshuler, D. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–33.

[2] Goldstein, D. (2009) Common genetic variation and human traits. *New England Journal of Medicine* **360**, 1696–1698.

[3] Lehner, B. (2011) Molecular mechanisms of epistasis within and between genes. *Trends in Gen* **27**, 323–331.

[4] Lewontin, R & Kojima, K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472.

[5] Slatkin, M. (2008) Linkage disequilibrium:understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**, 477–485.

[6] Lewontin, R. (1964) The interaction of selection and linkage. *Genetics* **49**, 49–67.

[7] Hill, W & Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theory of Applied Genetics* **38**, 226–231.

[8] Rogers, A & Huff, C. (2009) Linkage disequilibrium between loci with unknown phase. *Genetics* **182**, 839–844.

[9] Weir, B. (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* (Sinauer Associates, Sunderland, MA.), pp. 125–127.

[10] Zhang, Y, Zhang, J, & Liu, J. (2011) Block-based bayesian epistasis association mapping with application to wtccc type 1 diabetes data. *The Annals of Applied Statistics* **5**, 2052–2077.

[11] Watkins, W, Zenger, R, OBrien, E, Nyman, D, Eriksson, A, Renlund, M, & Jorde, L. (1994) Linkage disequilibrium patterns vary with chromosomal location: a case study from the von willebrand factor region. *American Journal of Human Genetics* **55**, 348–355.

[12] Jorde, L, Watkins, W, Carlson, M, Groden, J, Albertsen, H, Thliveris, A, & Leppert, M. (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *American Journal of Human Gen* **54**, 884–898.

[13] Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139–44.

[14] Dunning, A, Durocher, F, & Healey, C. (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67**, 1544–1554.

[15] Reich, D, Cargill, M, Bolk, S, Ireland, J, Sabeti, P, Richter, D, Lavery, T, Kouyoumjian, R, Farhadian, S, Ward, R, & Lander, E. (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

[16] Daly, M, Rioux, J, Schaffner, S, Hudson, T, & Lander, E. (2001) High-resolution haplotype structure in the human genome. *Nature* **29**, 229–232.

[17] Gabriel, S, Schaffner, S, Nguyen, H, Moore, J, Roy, J, Blumenstiel, B, Higgins, J, DeFelice, M, Lochner, A, Faggart, M, Liu-Cordero, S, Rotimi, C, Adeyemo, A, Cooper, R, Ward, R, Lander, E, Daly, M, & Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

[18] Carlson, C, Eberle, M, Rieder, M, Yi, Q, Kruglyak, L, & Nickerson, D. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120.

[19] Wang, N, Akey, J, Zhang, K, Chakraborty, R, & Jin, L. (2002) Distribution of recombination crossovers and the origin of haplo-type blocks: the interplay of population history, recombina-tion, and mutation. *American Journal of Human Genetics* **71**, 1227–1234.

[20] Patil, N, Berno, A, Hinds, D, Barrett, W, Doshi, J, Hacker, C, Kautzer, C, Lee, D, Marjoribanks, C, McDonough, D, Nguyen, B, Norris, M, Sheehan, J, Shen, N, Stern, D, Stokowski, R, Thomas, D, Trulson, M, Vyas, K, Frazer, K, Fodor, S, & Cox, D. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome. *Science* **294**, 1719–1723.

[21] Zhang, K, Deng, M, Chen, T, Waterman, M, & Sun, F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* **99**, 7335–7339.

[22] Anderson, E & Novembre, J. (2003) Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics* **73**, 336–354.

[23] Jeffreys, A, Kauppi, L, & Neumann, R. (2001) Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature Genetics* **29**, 217–222.

[24] Cullen, M, Perfetto, S, Klitz, W, Nelson, G, & Carrington, M. (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *American Journal of Human Genetics* **71**, 759–776.

[25] Kauppi, L, Sajantila, A, & Jeffreys, A. . (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the mhc class ii region. *Human Molecular Genetics* **12**, 33–40.

[26] Koch, E, Ristroph, M, & Kirkpatrick, M. (2013) Long range linkage disequilibrium across the human genome. *PLoS One* **8**.

[27] Schmenger, C, Hoegel, J, Vogel, W, & Assum, G. (2005) Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the european population. *Human Genetics* **118**, 276–286.

[28] Garrigan, D, Mobasher, Z, Kingan, S, Wilder, J, & Hammer, M. (2005) Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. *Genetics* **170**, 1849–1856.

[29] Devlin, B & Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**, 997–1004.

[30] Pritchard, J & et al. (2000) Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

[31] Price, A, Patterson, N, Plenge, R, Weinblatt, M, Shadick, N, & et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

[32] Consortium., I. H. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.

[33] Consortium., I. H. (2007) A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 851–861.

[34] Kaiser, J. (2008) Dna sequencing: A plan to capture human diversity in 1000 genomes. *Science* **319**, 395.

[35] Ruggieri, E & Lawrence, C. (2012) On efficient calculations for bayesian variable selection. *Journal Computational Statistic and Data Analysis* **56**, 1319–1332.

[36] Durbin, R, Eddy, S, Krogh, A, & Mitchison, G. (1998) *Biological Sequence analysis: Probabilistic models of proteins and nucleic acids*. (Cambridge University Press), p. Chapters 9 & 10.

[37] Chomsky, N. (1957) *Syntactic Structures*. (Mouton & Co.).

[38] Manning, C & Schutze, H. (1999) *Foundations of Statistical Natural Language Processing.* (MIT Press, Cambridge, MA).

[39] Reeder, J, Steffen, P, & Giegerich, R. . (2005) Effective ambiguity checking in biosequence analysis. *BMC Bioinformatics* **6**, 153.

[40] Ding, Y & Lawrence, C. (2003) A statistical sampling algorithm for rna secondary structure prediction. *Nucleic Acids Research* **31**, 7280–7301.

[41] McCaskill, J. (1990) The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers* **29**, 1105–1119.

[42] McKeown, M. (1992) Alternative mrna splicing. *Annual Review of Cell Biology* **8**, 133–155.

[43] Peltz, S & Jacobson, A. (1992) mrna strability:in trans-it. *Current Opinions in Cell Biology* **4**, 979–983.

[44] Bonhoeffer, S, McCaskill, J, Stadler, P, & Schuster, P. (1993) Rna multistructure landscapes. a study based on temperature-dependent partition-functions. *European Biophysics Journal* **22**, 13–24.

[45] Christoffersen, R, McSwiggen, J, & Konings, D. (1994) Application of computational technologies to ribozyme biotechnology products. *Journal of Molecular Structure* **311**, 273–284.

[46] Weidner, H, Yuan, R, & Crothers, D. (1977) Does 5s rna function by a switch between two secondary structures? *Nature* **266**, 193–194.

[47] Altuvia, S, Kornitzer, D, Teff, D, & Oppenheim, A. (1989) Alternative mrna structures of the ciii gene of bacteriophage l determine the rate of its translation initiation. *Journal of Molecular Biology* **210**, 265–280.

[48] Stormo. (1986) *Maximizing Gene Expression.*, eds. Reznikoff, W & Golg, L. (Butterworth Publishers, Stoneham, MA), pp. 195–224.

[49] Krogh, A, Brown, M, Mian, I, Sjolander, K, & Haussler, D. (1993) Hidden markov models in computational biology. *Journal of Molecular Biology* **235**, 1501–1531.

[50] Sakakibara, Y, Brown, M, Hughey, R, Mian, I, Sjlander, K, Underwood, R, & Haussler, D. (1994) Stochastic context-free grammers for trna modeling. *Nucleic Acids Research* **22**, 5112–5120.

[51] Lefebvre, F. (1996) A grammar-based uni

cation of several alignment and folding algorithms. *Proc Int Conf Intell Syst Mol Biol* **4**, 143–154.

[52] Dowell, R & Eddy, S. (2004) Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC Bioinformatics* **5**, 71–85.

[53] Knudsen, B & Hein, J. (1999) Rna secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**, 446–454.

[54] Do, C, Woods, D, & Batzoglou, S. (2006) Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics* **22**, 90–98.

[55] Mathews, D, Andre, T, Kim, J, Turner, D, & Zuker, M. (1998) *Molecular Modeling of Nucleic Acids*, eds. Leontis, N & SantaLucia, J. J. (American Chemical Society), pp. 246–257.

[56] Hofacker, I, Fontana, W, Stadler, P, Bonhoeffer, S, Tacker, M, & Schuster, P. (1994) Fast folding and comparison of rna secondary structures. *Montash. Chem.* **125**, 167–188.

[57] Carvalho, L & Lawrence, C. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *PNAS* **105**, 3209–3214.

[58] Lari, K & Young, S. (1990) The estimation of stochastic context-free grammars unsing the inside-outside algorithm. *Computer Speech and Language* **4**, 35–56.

[59] Hopcroft, J, Motwani, R, & Ullman, J. (2001) *Introduction to Automata Theory, Languages and Computation*. (Addison-Wesley Longman Publishing Co., Inc.), 2nd edition.

[60] Patterson, N, Price, A, & Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, 2074–2093.

[61] Goddard, K, Hopkins, P, Hall, J, & Witte, J. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *American Journal of Human Genetics* **66**, 216–234.

[62] Marth, G, Czabarka, E, Murvai, J, & Sherry, S. (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.

[63] Weinberg, C & Shi, M. (2009) The genetics of preterm birth: Using what we know to design better association studies. *American Journal of Epidemiology* **170**, 1373–1381.

[64] Uzun, A, Dewan, A, Istrail, S, & Padbury, J. (2013) Pathway-based genetic analysis of preterm birth. *Genomics* **101**, 163–170.

[65] Steffen, K, Cooper, M, & Shi, M. (2007) Maternal and fetal variation in genes of cholesterol metabolism is associated with preterm delivery. *Journal of Perinatal Medicine* **27**, 672–680.

[66] Mailman, M, Feolo, M, Jin, Y, Kimura, M, Tryka, K, Bagoutdinov, R, Hao, L, Kiang, A, Paschall, J, Phan, L, Popova, N, Pretel, S, Ziyabari, L, Lee, M, Shao, Y, Wang, Z, Sirotkin, K, Ward, M, Kholodov, M, Zbicz, K, Beck, J, Kimelman, M, Shevelev, S, Preuss, D, Yaschenko, E, Graeff, A, Ostell, J, & Sherry, S. (2007) The ncbi dbgap database of genotypes and phenotypes. *Nature Genetics* **39**, 1181–1186.

[67] Boyd, H, Poulsen, G, Wohlfahrt, J, Murray, J, Feenstra, B, & Melbye, M. (2009) Maternal contributions to preterm delivery. *American Journal of Epidemiology* **170**, 1358–1364.

[68] Ramamurthy, B, Jones, A, & Larsson, L. (2003) Glutathione reverses early effects of glycation on myosin function. *American Journal of Physiology* **285**, 419–424.

[69] Risch, N & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1616–1617.

[70] Li, H, Wei, Z, & Maris, J. (2010) A hidden markov random field model for genome-wide association studies. *Biostatistics* **11**, 139–150.

[71] Schaid, D, Rowland, C, Tines, D, Jacobson, R, & Poland, G. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

[72] Lin, D & Huang, B. (2007) On the use of inferred haplotypes in downstream analyses. *American Journal of Human Genetics* **80**, 577–579.

[73] Huang, B, Amos, C, & Lin, D. (2007) Detecting haplotype effects in genomewide association studies. *Genetic Epidemiology* **31**, 803–812.

[74] BROWNING, B & BROWNING, S. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology* **31**, 365–375.

[75] Cardon, L & Abecasis, G. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genetics* **19**, 135–140.