On populations, haplotypes and genome sequencing

by

Pierre Franquin

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

Courant Institute of Mathematical Sciences

New York University

September 2012

Bhubaneswar Mishra — Advisor

Dedication

To Emily and my family.

Acknowledgments

The present work would not have came to a reality without the help, support and dedication, directly or indirectly, of many great people.

First and foremost, my advisor, Professor Bud Mishra, without whom there would be no story. It has been an honor and a privilege to see a great mind at work and I have learned many things from him. He guided me throughout my Ph.D. and has been a great scientific mentor, an inextinguishable source of ideas and something that never hurts, an amazing human being full of humor. As a student, being able to work on a project that might have a significant impact in the life of millions is both overwhelming and exciting and I am really grateful to have been a part of that. He definitely is the major contributor of this thesis.

From a scientific and academic perspectives, I would also like to thank the members of my committee, Professors Frank Hoppensteadt, Raul Rabadan, Mickey Atwal and Ernest Davis. They have helped me understand more in depth some problems faced, especially in population genetics. They have also been a great aid in making this document coherent.

A Ph.D. is definitely more a marathon than a sprint. During the five years of work, you will inevitably have ups and downs. And in the midst of the really bad

times, having a constant figure really helps get passed the storm. I want to give a special thanks to my wife Emily who has always been there for me. She has helped me overcome the worst difficulties. She has been supportive, loving and caring. She should also be thanked for her work of edition on this document. If you are not reading some frenglish text, it is all due to her.

I want to thank my family and friends who are in France. When I told my family I was leaving for the United States, I never felt the slightest doubt or reluctance in their minds about it. They have always told me to pursue my dreams and have been extremely supportive. As with my wife, my family and friends have been my rock for the past five years, the people I could count on and who always gave me confidence.

Finally, I want to thank my family-in-law. I arrived without knowing anybody in this country and they have made me feel that I also had a family in the US. Their kindness and love for me made me feel less on my own and therefore have allowed me to focus on my work.

Abstract

Population genetics has seen a renewed interest since the completion of the human genome project. With the availability of rapidly growing volumes of genomic data, the scientific and medical communities have been optimistic that better understanding of human diseases as well as their treatment were imminent. Many population genomic models and association studies have been designed (or redesigned) to address these problems. For instance, the genome-wide association studies (GWAS) had raised hopes for finding disease markers, personalized medicine and rational drug design. Yet, as of today, they have not yielded results that live up to their promise and have only led to a frustrating disappointment.

Intrigued, but not deterred by these challenges, this dissertation visits the different aspects of these problems. In the first part, we will review the different models and theories of population genetics that are now challenged. We will propose our own implementation of a model to test different hypotheses. This effort will hopefully help us in understanding whether the research community expectations were unreasonably too high or if we had ignored a crucial piece of information.

When discussing association studies, we must not forget that we rely on data that are produced by sequencing technologies, so far available. We have to ensure that the quality of this data is reasonably good for GWAS. Unfortunately, as the reader will see in the second part, despite the existence of a diverse set of sequencing technologies, none of them can produce haplotypes with phasing, which appears to be the most important type of sequence data needed for association studies. To address this challenge, I

propose a novel approach for a sequencing technology, called SMASH that allows us to create the quality and type of haplotypic genome sequences necessary for efficient population genetics.

Contents

Dedication iii

Acknowledgments iv

Abstract vi

List of Figures xii

List of Tables xiv

List of Appendices xvi

Introduction 1

1 Population Genetics 11

1.1 Models 11

1.1.1 Wright-Fisher 11

1.1.2 Moran 17

1.1.3 Coalescence 20

1.2 Making sense out of sequence? 23

1.2.1 Single Nucleotide Polymorphisms 23

1.2.2 Linkage Disequilibrium 25

2 Simulations **28**

2.1 Implementation of our Model 28

2.2 Common Disease Common Variant 32

2.2.1 Theory 32

2.2.2 Debate 34

2.3 Simulations 38

3 Genome Wide Association Study **43**

3.1 Status of GWAS 43

3.2 HapMap 51

3.3 Haplotype: The Missing Link? 53

4 Sequencing Technologies **57**

4.1 Technologies 57

4.1.1 Sequencing 57

4.1.2 Mapping 61

4.2 Assemblers 62

4.2.1 Phrap 62
4.2.2 TIGR 63
4.2.3 CAP3 64
4.2.4 Celera 66
4.2.5 Arachne 67
4.2.6 EULER 68
4.2.7 SOAPdenovo 69
4.2.8 AllPaths 70
4.2.9 Abyss 72
4.2.10 SUTTA 72

5 SMASH **74**

5.1 Sequencing Technology 76
5.1.1 Optical Restriction Fragments Mapping 77
5.1.2 Optical Probes Mapping 80
5.1.3 Results 83
5.2 Assembler Algorithm 86
5.2.1 Results 91
5.2.2 Complications 96

5.3 Improvements 102

5.3.1 Design of gapped probes 102

Conclusion **106**

Appendices **110**

Bibliography **119**

List of Figures

2.1 Constant size population of 5000 individuals with no mutations after the 200th generation 39

2.2 Constant size population of 5000 individuals with mutations . . . 40

2.3 Constant size population of 1000 individuals with mutations, following ten SNPs 41

5.1 880 bp fragment resolved using 4% PAGE gel. 83

5.2 Overlayed fluorescent images of labmda DNA molecules. 84

5.3 Experiments with E. coli K-12 genome. 85

5.4 Noise Model. 87

5.5 Branch and Bound Algorithm. 90

5.6 Sequencing errors per 10kb sequence for solid (no universal bases) probes 92

5.7 Sequencing errors per 10kb sequence for gapped probes 93

5.8 Percentage of correct assembly of our sequence for different probe patterns. 102

9 Graph Construction to Prove that SMASH-P is NP-Complete . . 114

10 Fast Bottleneck and Fast Population Growth 116
11 Slow Bottleneck and Fast Population Growth 117
12 Slow Bottleneck and Slow Population Growth 118

List of Tables

2.1 Running time of simulations with different parameters. u is the mutation rate per generation per sequence and r is the recombination rate per generation per sequence. The simulations were run on a 3.06 GHz Intel Core 2 Duo with 4 GB of RAM. The code is written in python and interpreted using pypy. 42

5.1 Percentage of sequence correctly assembled for different values of false negatives while other parameters (false positives, window error size, probe pattern) vary 94

5.2 Percentage of sequence correctly assembled for different values of false positives while other parameters (false negatives, window error size, probe pattern) vary 95

5.3 Percentage of sequence correctly assembled for different values of sizing errors while other parameters (false negatives, false positives, probe pattern) vary 96

5.4 Percentage of sequence correctly assembled for different probe patterns while other parameters (false negatives, false positives, window error size) vary 97
5.5 Coverage of position i by ungapped probes of size 4 103
5.6 Coverage of position i by two different gapped probes of size 4 with two universal bases 103
5.7 Value of the spectral gap for the different 6-mers 105

List of Appendices

Appendix A **110**

Appendix B **112**

Appendix C **116**

Introduction

This thesis addresses two important problems in today's computational biology. First is the problem of population genetics and more precisely, how to make sense of genomic sequences. The large amount of data containing useful genetic information could someday allow us to treat diseases and develop personalized medicine, which is one of the most exciting challenges of the new century. However, since the human genome was first fully sequenced, few usable results have been found. Different strategies have been proposed but none of them has been reliable enough to provide a breakthrough in genetic therapy. This raises the question of where the obstacles to genetic medicine may lie. Have the hypotheses, theories and models proposed been wrong or at least incomplete or might the problems encountered be due to the quality of the data on which we are currently running experiments?

This point will lead to the second topic addressed in this dissertation; namely sequence quality. An essential limiting factor for any population study is the quality of the sequences we want to use to run our models or test our theories which is why having quality sequences is an extremely important matter. As we approach the tenth anniversary of the completion of sequencing of first hu-

man genome, sequencing technologies have become less and less expensive, but it seems that the quality of the sequences we obtain from those technologies is not good enough to lead significant population studies. In the ten years since we first sequenced the human genome, the problem of sequencing the human genome has not have been fully solved. Genome sequencing is a very complex problem and it is interesting to see how it has been tackled and if we can improve it. You need to learn to walk before you can run, which in our case would imply that you need to have good quality data before you can study them.

In this introduction, I will present the different challenges posed by these two problems; how to make sense of genomic sequences and how to improve the quality of the sequences we get. I will also present the reasoning behind some solutions that this thesis will present for these problems.

Motivation

As stated earlier, being able to find alleles responsible for diseases, isolate them, understand their relationships to one another and finally propose an approach to treat those diseases are some of contemporary genetics' most exciting challenges. Why is sequencing so important? Sequencing a genome is more a tool than a goal in itself. Using this raw string of letters tool, we are able to compare sequences to each other and gather important data about human variation. On the other hand, it is futile to study genomic features without good genomic data. This is why the two problems addressed here are so intertwined. Sequencing genomes and studying them are important in many ways in biology and medicine. The study

of genomes has myriad applications. It allows us to have a better understanding of evolution by comparing genomes of different species. It can also allow us to understand regular traits or diseases by comparing the genomes of wild-type and mutants, patients and normals. Another goal of genomic study is to be able to automatically find regions of the genome that have a particular significance such as genes, splicing sites, regulatory regions, etc, greatly cutting the cost of those operations. We can also study the behavior of the genome as a whole and have a better understanding of intergenic regions.

This is a tiny list, far from exhaustive, of the different possibilities that those problems offer to solve and I cannot imagine the invaluable breakthrough there could be if only one of those problems could be better elucidated by the present thesis.

Contributions

The present thesis will contribute in two different ways.

1. A new approach and design to sequence whole genomes called Single Molecule Approach to Sequencing by Hybridization (SMASH). This design will allow us to sequence haplotypes in an inexpensive way. While relying on a technology known to be NP-complete, the combination of this approach and another technology allow us to tame the complexity of the problem. This work builds on the earlier unpublished work with Anantharaman, Lim, Reed and Mishra.

2. A Wright-Fisher model that will let us try different hypotheses about populations with common features such as mutations and recombinations but also more advanced ones such as scenario of population size fluctuation and type of mutations followed (e.g. lethal, selectively neutral, giving heterozygous advantage).

Outline

The first three chapters will address population genetics. While the first chapter is more of an introduction to the different existing population models and different characteristics of interest inside the genome, the second chapter will introduce the model developed in this thesis; namely, the common disease common variant (CDCV) hypothesis and what our model is capable of inferring. The third chapter will discuss whole genome association studies, their successes and problems, and will highlight what we think might be a reason why those studies have been more or less futile so far, given the unavailability of haplotype sequences. Knowing this, chapter four will present different sequencing technologies and algorithms associated with them. Finally, chapter five will introduce our new approach to sequencing, SMASH and will show why this might be a major advance in the field of sequencing.

Motivations

The main idea motivating this thesis is to find a way to design effective genome-wide association studies (GWAS). We wished to address various problems hindering association studies that have been performed to date with disappointing results. Some simple models such as the common disease common variant hypothesis was hoped to simplify the analysis, but CDCV hypothesis still remains controversial and does not work well for reasons that will be discussed in later chapters. We sought to use in silico/simulation based model to design effective GWAS.

The casecontrol design study has been the most widely applied strategy of association study for characterizing genetic contributions to disease. The advantages of this design are it can be done quickly with a large number of cases and controls and the cases can be efficiently genotyped and compared to the controls. On the other hand this approach is prone to bias, especially when it comes to selecting individual for the control population as the selection process often leads to associations that are due to some population stratification rather than a real association for the disease. Another type of design would be to use cohorts. This approach would allow the study to be significantly less biased but the resources

both financial and time-wise make cohort studies hard to implement. Yet another design is to use family-based controls. This strategy is totally immune to population stratification but is highly sensitive to genotyping errors and can be hard to implement in the case of a late onset disease. Another side effect of this approach is the loss of statistical power to detect genuine allelic association. The different designs of those GWAS also rely on assumption about the variants that are supposed to be found, such as SNPs and on the type of data that underlie these studies (e.g., data obtained through technologies that give genotypes assembled from short reads).

Results found using GWAS have been encouraging and disappointing in the same time for complex traits. While a set of SNPs can be found to be statistically significant, individually and in aggregate those SNPs seem to only account for a small proportion of genetic variance. Interpreting the low predictive power of the variants has been called the missing heritability problem. The obstacles encountered with the missing heritability problem overlap somewhat with the problems encountered in GWAS (difficulties defining phenotypes, population stratification, common variations that are left out such as CNV or gap in SNPs coverage).

The power of GWA studies can be greatly increased if augmented with the knowledge of haplotypes and more specifically, phased haplotypes. Associations that were impossible to detect without phased haplotypes could become detectable. Even more complicated phase-dependent interactions of variants in linkage equilibrium have also been suggested as possible causes of missing heritability. But the current true haplotypes cannot be obtained accurately and even with errors,

the cost, whether it is money for the experiment or the computational cost to phase the haplotypes, is largely prohibitive. For those reasons, the significance of phased haplotypes and the gain of statistical power of GWAS using them is yet to be determined but rare variations are now accepted as being an important actor in common as well as rare diseases.

Current studies ignore the phase of DNA. While some projects have included haplotype in their analyses, they have generally assessed linkage disequilibrium without directly examining the precise layout of genes on two homologous chromosomes. It is more difficult to sequence a human genome with phased haplotypes than it is to simply have the overall sequence, without worrying about the origin of said sequence; i.e. which one of the pair of homologous chromosomes that sequence belongs to. In some cases, however, it is crucial to understand which copy of the chromosome carries a particular variant.

Allele-specific expression, for example, is when the copy of a gene on one copy of a chromosome is expressed while the copy on the other chromosome (the trans copy) is suppressed. It has been estimated that 1-5% of human genes are affected by allele-specific expression.

One mechanism for differential expression of the two alleles is that a transcription factor binds preferentially to a sequence on one chromosome copy over another, due primarily to differences in sequence. These sequences are, in turn, heritable, therefore one parent can pass along an allele that will be more highly expressed than that of the other parent.

Another mechanism for allele specific expression is epigenetic changes or changes

that affect phenotype which come from a source other than the sequence of bases on a strand of DNA. Methylation of chromosomal regions is one common form of epigenetic suppression of genes. Some methylation patterns are based on which parent a chromosome comes from (i.e. for certain regions of the genome, an individual will always express the maternal copy, and for other regions, the paternal copy). Other methylation patterns, however, appear to be the result of interactions betweens single nucleotide polymorphisms (SNPs) that occur both within one copy of a chromosome (cis acting) and between SNPs on different chromosomes (trans-acting). In the case that a SNP interaction affects methylation patterns on an allele-specific basis, the result is known allele-specific methylation.It has been suggested that allele-specific methylation may play a role in type-2 diabetes.

Copy number of genes can play a role in expression, and knowledge of true hapotypes can help in understanding the effect of cis-acting copies of a gene, or portions of a gene. For example, one copy of a gene for which an individual is heterozygous may be amplified in a cancerous state. Understanding the effects of this gene may require understanding its effects on the cis strand, thus knowledge of the sequence of that same strand of DNA.

Compound heterozygosity is a term used to describe two homologous copies of a region that have unique variants, but the variants occur at different locations within that region. In the situation of compound heterozygosity, the combined effect of these variants is different than what would result from having the variants on one single copy of the region. Because of this, understanding compound

heterozygosity and determining the risk of an individual for a disease in which compound heterozygosity plays a role, requires assessment of haplotype phasing. Diseases affected by compound heterozygosity include cerebral palsy, a glycogen storage disorder, and hyperphenylalinemia, among others. Compound heterozygosity may also play a role in cancer, where the effect of a deleterious mutation in one copy of a chromosome is potentiated by a mutation in the same region, but at a different location in the homologous chromosome.

Phase information also appears to be important for population genetics studies, as it has been found that greater differentiation of populations, and thus resolution of differences between populations can be found when haplotypes are included in the study. Similarly, phase information can enhance studies examining evolutionary patterns.

To design efficient GWAS, the present thesis aimed to try out different GWAS designs and show which ones are the best. This process starts by creating an accurate data set using a coalescent or Wright-Fisher approach to model diseases and selection and by carefully implementing and designing good algorithms, data structures and optimizations scheme (such as parallelization) to do so. With this model in hand, we could then try different designs of GWAS based on genotypes, family trees, haplotypes, number of individuals, etc and assess their effectiveness. Assuming that the addition of phased haplotypes will give a significant boost to the power of those studies, we will then need to design an effective sequencing technology that will sequence whole haplotype genomes in order to feed real data to our population model. Finally, we would like to enable a realistic experiment

(e.g. Wellcome Trust Case Control).

Unfortunately, the scope of this project was too ambitious given the constraint of time and resources I had as a PhD student in a small bioinformatics laboratory. The following subset of goals has been achieved:

1. A new approach and design to sequence whole genomes called Single Molecule Approach to Sequencing by Hybridization (SMASH). This design will allow us to sequence haplotypes in an inexpensive way. While relying on a technology known to be NP-complete, the combination of this approach and another technology allow us to tame the complexity of the problem. This work builds on the earlier unpublished work with Anantharaman, Lim, Reed and Mishra.

2. A Wright-Fisher model that will let us test different hypotheses about populations (and their diseases) with common features such as mutations and recombinations but also more advanced ones such as scenario of population size fluctuation and type of mutations followed (e.g. lethal, selectively neutral, giving heterozygous advantage).

Chapter 1

Population Genetics

1.1 Models

1.1.1 Wright-Fisher

This model was found independently and almost simultaneously by Fisher [**?**] and Wright [**?**] although Fisher had come very close almost a decade earlier. Let us consider the simplest possible case, which is founded on several simplifying assumptions. We envision a diploid population of size N which could have been also modeled as a haploid population of size $2N$. Further, we assume discrete and non-overlapping generations. In other words, we assume that reproduction and death are simultaneous for all individuals within the population. This assumption, while appearing very unrealistic, does not affect the asymptotic properties in any substantial way. We are going to focus on the case where the population size is constant. Important values of our model will be different if we instead assume a fluctuating population size (growing, shrinking, or both). We are also

going to assume that all individuals are equally fit. It is convenient to study a simple model albeit unrealistic. Fortunately, this hypothesis can be relaxed easily. Similarly, we will assume no recombination or mutation, an assumption which can also be relaxed. Finally, the mating process within this population is assumed to be random (panmictic), in other words there is no population structure.

Thus, in this simplest form, the model does not permit any mutation or recombination and there is no selective force between two alleles A and a at the same locus. We are going to pay attention to the number X of A alleles (or genes). Obviously, $X \in \{0, 1, 2, ..., 2N\}$. At each generation g, the number X will be noted $X(g)$. To derive a generation $g+1$ from a generation g, each gene (allele) gives birth to some number of offspring (which are the exact copies of itself) and dies immediately after that, thus living only one generation. $X(g+1)$ is therefore a binomial random variable with index $2N$ and parameter (probability of success) $\frac{X(g)}{2N}$. So, if $X(g) = i$, the probability p_{ij} that $X(g+1) = j$ is given by:

$$p_{ij} = \binom{2N}{j} (\tfrac{i}{2N})^j (1 - \tfrac{i}{2N})^{2N-j}$$

Here we are studying a very simple definition of the model to see the effects of stochastic variations in gene frequencies without any complications but the model can, of course, be enriched by adding different mechanisms such as mutations or selection.

If we go back to our simple model, we can make few easy observations. Since

there is no mutation, the states $X = 0$ and $X = 2N$ are absorbing. Without the possibility of mutating, once an allele has disappeared, it cannot reappear (here, with $X = 0$, the population will have allele a) and conversely, once the allele is present in everybody, the next generation, being a sample of the current generation, has to have the allele (here, the population will have allele A). We could rephrase this statement by saying that, whatever the value of $X(0)$, eventually X will fall into one of the two absorbing states. That translates mathematically to $\lim_{g\to\infty} X(g) = 0$.

We can study the probability of absorption in such a model. The probability of extinction (when $X = 0$) given that there were initially i alleles A in our population can be seen as $\lim_{g\to\infty}(X(g) = 0|X(0) = i)$. An easy way to study the probability of absorption or fixation is to use the expectation value of our variable. The constancy of expectation gives us:

$$E(X(g)) = E[E(X(g)|X(g-1))] = E(X(g-1)) = ... = E(X(0)) = i$$

We know that $E(X(g)) = 0.u_{i,0} + 2N(1 - u_{i,0})$ where $u_{i,0}$ is the probability of going from i alleles to 0. Therefore, we have

$$0.u_{i,0} + 2N(1 - u_{i,0}) = i \Rightarrow u_{i,0} = 1 - \frac{i}{2N}$$

By following the same reasoning, we can compute the probability of fixation (when $X = 2N$) given that there were initially i alleles A in our population

$$0.(1 - u_{i,2N}) + 2N.u_{i,0}) = i \Rightarrow u_{i,2N} = \frac{i}{2N}$$

We could have thought of this value more literally by saying that eventually, every gene in the population is descended from one unique gene in the first generation. The probability that such a gene is A is simply the fraction of A genes in the initial population.

Another value of interest with this model is the mean time until absorption. It is a very complicated value to compute precisely and we will only compute an approximation of it. One value that is simple to compute is the mean time to absorption when there is just one allele of type A in our first generation. Starting in state $X(0) = 1$, we will compute the expected number of visits to a state j we make before we reach a state of absorption (0 or $2N$). We denote the mean number of generations to absorption in 0 or 2N , given that we started with one allele A as $\bar{t}_1$. We need to sum up the expected number of such visits for all j, avoiding states 0 and $2N$

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j}$$

where $\bar{t}_{1,j}$ is the mean number of times that the number of A alleles takes the value of j. Again, both Wright and Fisher found that $\bar{t}_{1,j} \approx \frac{2}{j}$. Since $\sum_{i=1}^{N} \frac{1}{i} = \log(N) + \gamma$ where γ is the Euler's constant, we find that

$$\bar{t}_1 = \sum_{j=1}^{2N-1} \bar{t}_{1,j} = \sum_{j=1}^{2N-1} \frac{2}{j} = 2 \sum_{j=1}^{2N-1} \frac{1}{j} = 2(\log(2N-1) + \gamma)$$

As stated earlier, we could find solutions for a more general i but in practice, simple expressions for those solutions have not yet been found and may never be found. We can, however, compute an approximation for $\bar{t}_i$. We use the first step analysis where we start from a state i and in the first step visit some intermediate state k. We define $M = 2N, i/M = x, k/M = x + \delta x$ and $\bar{t}_i = \bar{t}(x)$. We can write

$\bar{t}_i = \sum_{k=0}^{M} p_{ik}\bar{t}_k + 1$

as

$\bar{t}(x) = \sum P(x \to x + \delta x)\bar{t}(x + \delta x) + 1 = E(\bar{t}(x + \delta x)) + 1$

Now, by applying the Taylor's series to $\bar{t}(x)$ we have

$\bar{t}(x) \approx \bar{t}(x) + E(\delta x)\bar{t}(x)' + \frac{1}{2}E(\delta x)^2\bar{t}(x)'' + 1$

Using the fact that the expectation of the binomial random variable is $E(X) = np$, we have

$E(x + \delta x) = E(\frac{j}{M}) = \frac{E(j)}{M} = \frac{M \times \frac{i}{M}}{M} = \frac{i}{M}$

Since $x = \frac{i}{M}$ and $E(x) = x$, we can say that $E(\delta x) = 0$ and therefore $E(\delta x)\bar{t}(x)' = 0$. Now we will compute $E(\delta x)^2$. In our case, $E(\delta x)^2 = Var(\delta x)$

since $Var(\delta x) = E(\delta x)^2 - [E(\delta x)]^2$ and $[E(\delta x)]^2 = 0$ as just shown.

$$Var(x + \delta x) = Var(\tfrac{j}{2N}) = \tfrac{Var(j)}{4N^2} = \tfrac{2N\frac{i}{2N}(1-\frac{i}{2N})}{4N^2} = \tfrac{x(1-x)}{2N}$$

By plugging in this result to our expression of $\bar{t}(x)$, we have

$$\bar{t}(x) \approx \bar{t}(x) + \tfrac{1}{2}\tfrac{x(1-x)}{2N}\bar{t}(x)'' + 1$$
$$x(1-x)\bar{t}(x)'' \approx -4N$$

The solution to this equation, subject to the boundary conditions $\bar{t}(0) = \bar{t}(1) = 0$ is

$$\bar{t}(x) = -4N \int\int \frac{1}{x(1-x)}$$
$$\bar{t}(x) = -4N \int ln(x) + ln(1-x)$$
$$\bar{t}(x) = -4N((xln(x) - x) + ((1-x)ln(1-x) - (1-x)))$$
$$\bar{t}(x) = \approx -4N(x \log x + (1-x)\log(1-x))$$

This computation is called the diffusion approximation to the mean absorption time. If we initially start with one allele A which is equivalent to $x = \frac{1}{2N}$, the mean time to absorption is $\bar{t}(x) \approx 2 + 2\log(2N)$. If we started with a population with as many genes A as a, that is with $x = \frac{1}{2}$, the mean time is $\bar{t}(x) \approx 2.8N$ which is clearly longer than for the case with one mutant.

1.1.2 Moran

The main difference between the Wright-Fisher model we have just described and the Moran model [?] is the fact that we allow overlapping generations in the Moran model. Here, an individual is chosen randomly to reproduce and another one is chosen to die (it could be the parent chosen to reproduce but not the offspring). The offspring now lives in a population belonging to his parent's generation. Because it does not make much sense to talk about generations in this model, each time an individual is chosen to reproduce and another one to die, it will increment a variable t. This process of choosing a reproducing and a dying individual is called a birth and death process. We will consider a population of $2N$ haploids who could have either the allele A or the allele a and, as in the Wright-Fisher model, ignore selection or mutation.

We define X to be a random variable which represents the number of times the allele A is present within the population. It is of interest to calculate transition probabilities for the implied Markov chain. Suppose that in a population at a time t, which corresponds to the state X_t in the underlying Markov chain, the number of times allele A is present is i. Then, at time $t+1$, the number of copies of allele A can be either $j = i + 1$ if an individual with allele A is chosen to reproduce and an individual with allele a is chosen to die, $j = i - 1$ if an individual with allele a is chosen to reproduce and an individual with allele A is chosen to die or $j = i$ if an individual with allele A (resp. a) is chosen to reproduce and an individual with allele A (resp. a) is chosen to die. The probability of going from i to $i + 1$ is

$$p_{i,i+1} = \frac{i}{2N} \times \frac{2N-i}{2N}$$

With a similar reasoning, the probability of going from i to $i-1$ is

$$p_{i,i-1} = \frac{i}{2N} \times \frac{2N-i}{2N}$$

And the probability of staying with i A alleles is

$$p_{i,i} = (\frac{i}{2N} \times \frac{i}{2N}) + (\frac{2N-i}{2N} \times \frac{2N-i}{2N}) = \frac{i^2+(2N-i)^2}{4N^2}$$

Those transition probabilities can define a matrix which is a continuant since $p_{i,j} = 0$ if $|i-j| > 1$. Therefore, we can use the theory on continuant matrix to explicitly find the probability of fixation and the mean time of absorption. We can use concepts from the processes of birth and death to calculate these quantities.The birth and death process is a special case of the continuous time Markov process where the states represent the current size of a population and where the transitions are limited to births and deaths. When a birth occurs, the state goes from i to $i+1$ defined by the birth rate $\lambda_i = p_{i,i+1}$. Similarly, when a death occurs, the state goes from i to $i-1$ defined by the death rate $\mu_i = p_{i,i-1}$. We define $\rho_i = \frac{\lambda_1 \times \lambda_2 \times ... \times \lambda_i}{\mu_1 \times \mu_2 \times ... \times \mu_i}$. Since $\lambda_i = \mu_i$ in the Moran model, we have $\rho_i = 1$. Hence, the probability of absorption, whether it is is state 0 or $2N$ is

$$u_i = \frac{\sum_{k=0}^{i-1} \rho_k}{\sum_{k=0}^{2N-1} \rho_k}$$

$u_i = \frac{i}{2N}$

In the same fashion, we can compute the mean time to absorption. We can calculate the mean number of times the system is in a state j given that it started in a state i as

$\bar{t}_{i,j} = \frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j}$ for $j = 1, ..., i$

$\bar{t}_{i,j} = \frac{(1-\frac{i}{2N})\sum_{k=0}^{j-1}1}{1\times\frac{2N-j}{2N}\times\frac{j}{2N}}$

$\bar{t}_{i,j} = \frac{(2N-i)\times 2N}{2N-j}$

And for $j = i+1, ..., 2N-1$

$\bar{t}_{i,j} = \frac{u_i\sum_{k=j}^{2N-1}\rho_k}{\rho_j\lambda_j}$

$\bar{t}_{i,j} = \frac{(\frac{i}{2N})\sum_{k=j}^{2N-1}1}{1\times\frac{2N-j}{2N}\times\frac{j}{2N}}$

$\bar{t}_{i,j} = \frac{i\times 2N}{2N-j}$

Combining those two results, we can now compute the mean time to absorption

$\bar{t}_i = \sum_{j=1}^{2N-1}\bar{t}_{i,j}$

$\bar{t}_i = \sum_{j=1}^{i}(\frac{(1-u_i)\sum_{k=0}^{j-1}\rho_k}{\rho_{j-1}\mu_j}) + \sum_{j=i+1}^{2N-1}(\frac{u_i\sum_{k=j}^{2N-1}\rho_k}{\rho_j\lambda_j})$

$\bar{t}_i = \sum_{j=1}^{i}\frac{(2N-i)\times 2N}{2N-j} + \sum_{j=i+1}^{2N-1}\frac{i\times 2N}{2N-j}$

$\bar{t}_i = (2N-i)2N\sum_{j=1}^{i}\frac{1}{2N-j} + 2Ni\sum_{j=i+1}^{2N-1}\frac{1}{j}$

The fact that the Wright-Fisher model works generation-by-generation makes it an efficient model for computer scientists. It is easier to code and consumes less computational resources than the Moran model. On the other hand, the mathematical computations one may wish to perform are easier and more exact with the Moran model than the Wright-Fisher. For example, while we had to find an approximation for the mean time to absorption in the Wright-Fisher model, the computation is relatively simple and exact with the Moran model.

1.1.3 Coalescence

Both the Wright-Fisher model and the Moran model look forward in time. They try to predict which alleles will eventually fix or become extinct and how long it will take. The coalescence looks backward in time. The first to introduce the idea of following a pair of genes back to their common ancestor is Gustave Malécot [**?**] in 1942. Coalescence examines a concept known as time to most recent common ancestor (TMRCA). This answers the question; if we pick two genes from a Wright-Fisher population, how long has it been on average since the two genes departed from their most recent common ancestor (MRCA)? Instead of making a predictive statement as with the previous models, we are now making a historical statement. In 1966, Harris [**?**] and Lewontin and Hubby [**?**] extended the question to samples larger than two. Let's say we pick four genes from a Wright-Fisher population. We can ask the same question, how long ago did the genes in the sample share a common ancestor? Alternatively, we could ask; how many samples do we need to be reasonably sure of sampling the MRCA of the

entire population? The work of Ewens [?] and Watterson [?] were also stepping stones for the coalescent theory. In 1982, Kingman [?], [?] and [?] finally proved the existence of the coalescent process and showed that the n-coalescent or the coalescent for the sample of n genes holds for a wide variety of populations.

What is the probability for two genes to have a common ancestor j generations back in time? First, the probability that two genes choose the same parent the previous generation is $\frac{1}{2N}$ for a population made of $2N$ individuals. The first one choose freely but the second one has to choose the same parent. Therefore, the probability that two genes have a common ancestor j generations back in time is $(1-\frac{1}{2N})^{j-1}\frac{1}{2N}$ since samples from different generations are independent of each other. With the same reasoning, we can compute the probability for k genes to find a common ancestor. Actually, it is easy to compute the probability that k genes have k different parents (no coalescence event). The first can choose freely, then the second has to choose a different parent within a pool of $2N-1$ individuals, the third can only choose among $2N-2$ individuals and so on. It gives us

$$\frac{2N-1}{2N}\frac{2N-2}{2N}...\frac{2N-k+1}{2N} = \prod_{i=1}^{k-1}(1-\frac{i}{2N}) = 1-\sum_{i=1}^{k-1}\frac{i}{2N}+o(\frac{1}{N^2}) = 1-\binom{k}{2}\frac{1}{2N}+o(\frac{1}{N^2})$$

Here the $o(\frac{1}{N^2})$ is negligible since n is much smaller than N. This approximation means that we may discard the possibility for two pairs of genes to find a common ancestor in the same generation. So, with n much smaller than N, the probability that none of the k genes found a common ancestor is

$$1 - \binom{k}{2}\frac{1}{2N}$$

And therefore, the probability of a coalescent event to occur is

$$\binom{k}{2}\frac{1}{2N}$$

It is now easy to see the probability that two genes out of k find a common ancestor j generations ago is

$$P(T_k^j) \approx (1 - \binom{k}{2}\frac{1}{2N})^{j-1}\binom{k}{2}\frac{1}{2N}$$

Here, the time was discrete. We can easily change to a continuous time process. Usually, the scale is made so that one unit of time is equivalent to the average time for two genes to coalesce (which is $2N$ generations as shown above). Let $t = \frac{j}{2N}$ where j is the time measured in generations. The waiting time T_k for k genes to have $k-1$ ancestors in the continuous representation is exponentially distributed, $T_k \sim \exp(\binom{k}{2})$ and so

$$P(T_k \leq t) = 1 - e^{\binom{k}{2}t}$$

Here, we have given an introduction to different models of evolution of populations. We have made rather strong assumptions that do not reflect the reality of life. Obviously, those models have been studied deeply and developed further.

Concepts of biological relevance have been added, 'such as mutation, recombination, selection, linkage disequilibrium, population size fluctuations, population structure and so on in order to reflect a more realistic view of life. The model we have developed and that will be presented later in this dissertation will have those features implemented but here, we are looking at some paradigms we can find in population genetics models and a glimpse of the questions they can answer.

1.2 Making sense out of sequence?

One of the goals of population genetics is to explain the role of variations within the sequence in order to explain cause, prevalence and nature of human diseases. The idea is to connect the variations observed in sequences with different phenotypes. There are different type of variations such as insertions and deletions, mini and micro-satellites. The most common variation is the single nucleotide polymorphism (called SNP and pronounced snip). Along with linkage disequilibrium, a concept discussed below, SNPs are often used to define haplotypes.

1.2.1 Single Nucleotide Polymorphisms

Polymorphism in a sequence differs from a mutation only by an arbitrary definition. Often, if the variations (the different alleles) within the general population are observed at a frequency bigger than 1%, they are called polymorphisms while if the occur at a lower frequency, we refer to them as mutations. The most common polymorphism is the SNP. Most of the time, the SNPs have two different alleles, one major (more frequent) and one minor (less frequent). The most

common type of SNPs are transitions where purines are replaced by purines and pyrimidines by pyrimidines (T to C or A to G). Because the human is diploid, for each locus of a SNP, the individual can be either homozygous for the major allele (AA), heterozygous (Aa) or homozygous for the minor allele (aa). C to T SNPs are the most common in the human genome.

There are many ways to discover SNPs. When high throughput data became available, SNPs were discovered by aligning different clone overlaps of genomic DNA [?] and cDNA sequences [?], [?] and [?] or by reduced representation shotgun sequencing [?] and [?]. A vast number of SNPs have been detected with these technologies but their characteristics (allele, genotype and population frequencies) could not be determined by these strategies alone. Another problem was many of the SNPs identified by these methods could not be validated using an alternative method [?] or a different population [?]. This inconsistency suggests either that they are rare variants or that they are artifacts from the sequencing or cloning technologies.

As we have seen, a lot of effort has been spent on identifying and characterizing SNPs. Their abundance in the genome is thought to make them the perfect target in the construction of very high resolution genetic maps in humans. However, when it comes to disease association studies, this abundance does not necessarily guarantee an accurate detection of causal genetic variants. There is a high degree of correlation among SNPs which makes it hard to determine which of the SNPs are causal for the phenotype associated with it [?], [?]. Still, the possibility of being genotyped in a large scale has put their discovery and characterization

high among the priorities of the Human Genome Project [?].

About 9 million SNPs have been discovered so far. In the human genome, SNPs occur roughly every 200 base pairs [?]. Some of those SNPs are rare, meaning they occur only once (singletons) or twice (doubletons) in a human population sample consisting of several hundred people. SNPs that occur more often ($\geq 5\%$) are described as common variants. Those common SNPs are at the heart of one of the most controversial theories in contemporary population genetics: the common disease-common variant hypothesis (CDCV) [?], [?]. We will study this theory and test it later in this document.

1.2.2 Linkage Disequilibrium

As we have seen with SNPs, their study alone is not quite sufficient to detect the precise location of alleles responsible for phenotypes. We know that alleles on the genome are not independent of each other and therefore, studying their non-random association might be crucial to have a better understanding of regular phenotypes or diseases. This type of association between two or more loci is referred to as linkage disequilibrium (LD). The first to introduce this terminology were Kojima and Lewontin [?]. Linkage disequilibrium is very important because it affects and is affected by many factors. Using LD, we can get information about past events like recombination but also have a better idea of the breeding system, population divisions and histories. The study of linkage disequilibrium might also shed light on selection.

There are many definitions of LD but they all rely on the same quantity D.

Between alleles at two loci, D_{AB} is defined as

$D_{AB} = p_{AB} - p_A p_B$

with p_{AB} being the frequency of gametes carrying the pair of alleles A and B at two loci and p_A, p_B being the frequencies of those alleles. At first, the term gamete was used for allowing the loci to be on different chromosomes but the most common application now is with two loci on the same chromosome. In this case, the pair AB is called a haplotype and p_{AB} refers to the haplotype frequency. We will see the importance of haplotype later in this document. As stated earlier, many different definitions of LD are used. Lewontin defined D' to be the ratio of D to its maximum possible absolute value [**?**]. $D' = 1$ when one of the four haplotypes is absent, regardless of the haplotype frequency. Another value that is often used is $r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$. r^2 is a correlation coefficient of 1 or 0 (all or nothing) indicator variables indicating the presence of A and B. To try if a particular allele is associated with a genetic disease, $\delta_A = p_A + \frac{D}{p_B}$ is used [**?**]. If $D = 0$, there is linkage equilibrium (LE) which means alleles are independent of one another. Linkage equilibrium has some similarities with the Hardy-Weinberg equilibrium (HWE) in that they both imply that alleles at different loci are randomly associated. The HWE is established in one generation of random mating, whatever the gene frequencies were. If there is a shift from HWE, this means something unusual might be happening, whether it is selection, inbreeding or just genotyping mistakes. LE and HWE also differ from each other because HWE needs just one generation to be set while D decreases depending on the recombination frequency between two loci. Anyhow, LE will be reached,

but usually after multiple generations unless some other factors intervene such as selection, gene flow, mutation or genetic drift. This is why it is so interesting to study linkage disequilibrium because it gives some insight about past events.

Chapter 2

Simulations

2.1 Implementation of our Model

As introduced in the first chapter, there are two main paradigms for population models, backward or forward in time, coalescence illustrating the former and Wright-Fisher the later. The Wright-Fisher model might seem better suited since it is not an approximation as the coalescence is. Yet the coalescence model is more widely used since the running time of the Wright-Fisher model is prohibitive as soon as some parameters become too large (e.g. population size or number of generations). On the other hand, with the amount of data available growing at a quick pace, analyzing data with the coalescence might be more problematic than desired and no consensus method has yet been chosen between rejection algorithms, importance sampling, Markov chain Monte Carlo or approximate bayesian computation. Anyhow, a choice has to be made. For our model we selected the Wright-Fisher model.

As stated earlier, the problem with the Wright-Fisher implementation is its computational cost. There is a twist on the implementation of this model described in [**?**] that speeds up the process greatly. The classic implementation of the model simulates the genealogy generation by generation. In the accelerated implementation, we look at the genealogy for a few generations ahead and treat only those individuals whose genetic material will participate in future generations. In the current generation, we simulate a genealogy for the next k generations but we do not create the individuals of the next generation. We can now detect if a chromosome has undergone recombination or not during those generations. If it recombines, we check if any of the descendants that has part of its material will survive at generation k. If it does not recombine, we check if the genetic material is saved in the kth generation. If none of those two conditions are true, then we know that the genetic material of this individual will be lost by the kth generation at most and it is therefore useless to simulate him. This heuristic allows to simulate only the individuals that will participate in the genetic material present in the future, thus cutting down a lot of useless computation. The new generation is created based only on the people that will contribute later on. Then further genealogy is updated by one generation, and the process is repeated as long as the number of generations defined in the parameters is not reached.

This strategy looks good and promising, unfortunately there are important caveats to this Marjoram's acceleration. First and foremost, we have to simulate a genealogy to determine which individuals contribute to the genetic information of future generations. This works as long as the genealogy itself does not depend

on who is picked. To study rudimentary features (e.g. mutations, recombinations, islands of population), this approach is perfect but if we want to study more complex features such as lethal mutations or heterozygous advantage, one faces a problem. For example, with heterozygous advantage, an individual does not have the same chance of getting selected for the next generation depending on his genetic material. Therefore, the genealogy of the population is not random anymore and one can no longer randomly create a genealogy and see who is participating in the future. Another problem with this approach is the gain of time that is claimed. The code that is given with Marjoram's paper is erroneous and does not do what it claims. It is therefore hard to assess the real advantage gained by this implementation. We spend most of our time either recombining or mutating genomes. A careful implementation of those features is important so as not to waste time unnecessarily. The difference of time between the regular model and the accelerated version is about a factor of ten which is much larger than the factor of 5 that would be at most expected. This is primarily because there was something wrong in the implementation of the accelerated version. At the end, not being able to implement features for complex diseases and a gain of time that is not that impressive, we decided to carefully implement a regular version of the Wright-Fisher model. In addition to the regular mutation and recombination, we have also implemented more realistic features. On top of the mutation and recombination rates, we introduce the possibility of changing the size of the population. We can make it grow or shrink depending on a growth function. We can define this function on a certain number of generations. This

approach lets us simulate rapid growth over few generations or a population bottleneck effect on the results of the simulation. We can also combine shrinkage and growth of the population as many times as we want. Basically, any scenario one may want to define in term of population size fluctuation is possible. Another feature is the presence of recombination hotspots that can also be seen as regions of linkage disequilibrium. We can define regions of high recombination rate where recombinations can occur easily and regions of small recombination rate where few recombinations will occur (thus having a strong linkage disequilibrium). This technique mimics the linkage disequilibrium and haplotype block structure of the genome. We also want to simulate different types of diseases, from Mendelian to complex ones. We allow the user to define list of SNPs for which, if a mutation occurs, the individual will die. In this model, rare diseases are influenced by strong selection where the mutation is lethal before an individual can have any offspring. On the other hand, we can define another set of SNPs that would make the individual sick but would not prevent him from propagating those mutations. This category represents the case of mutations that are selectively neutral and are typically mutations that will define common diseases such as cancer or diabetes. We also implemented a heterozygous advantage feature. It is known that at some loci, while having two mutant alleles might lead to grave problems, being a heterozygote might give an advantage against some disease. A famous example is the thalassemia mutation that confers a certain protection against malaria but will cause blood disease if present in the two chromosomes. At the locus of interest, if an individual has no mutation, he will have a slightly lower probability

of being selected than if he had one mutation while his chances of being chosen if he has two mutations decrease dramatically. If we follow more than one locus, the effect will be cumulative. Let us say we follow i loci. The probability for this individual to be picked will be $\prod_i p_i$.

2.2 Common Disease Common Variant

2.2.1 Theory

The common disease common variant hypothesis was first proposed by Lander in 1996 [**?**]. This hypothesis says that the variants that are responsible for common diseases are reasonably frequent in the population (usually between 1 and 10%). For each of the loci responsible for the disease, there will be one or a few predominant disease alleles. This theory raised hopes for diagnosing and creating therapy against those diseases. Indeed, if only a handful of genes are responsible for a disease and that within this handful of genes, the allelic spectrum is simple (not very diverse), the resources needed to detect them would be reasonable. Lander and Reich later developed a model [**?**] to explain this theory.

The model is based on some assumptions. It starts with an panmictic ancestral population of fixed size ($N = 10000$). This population expands instantaneously to its current number ($N = 6 * 10^9$). The mutation rate μ is also defined as $3.2 * 10^{-6}$. Then, a few values are defined as follows:

- f is the total frequency of the set of disease alleles in the current population

- f_0 is the equilibrium frequency of the class of disease alleles (frequency expected under the balance between mutation and selection).

-f_{exp} is the frequency for the class of disease alleles just before the population expansion.

The probability that two alleles within the disease class are the same is: $\phi_{disease} = \frac{1}{1+4N\mu(1-f_0)}$ where N is the effective overall population size and μ is the probability that a non-disease allele will mutate into a disease one. The common variant/common disease hypothesis can now be expressed as the prediction that $\phi_{disease}$ is high for the disease loci responsible for most of the population risk for common diseases.

A rare disease will have a low f_0 and a common disease a high one. Since the mutation rate is constant and f_0 is determined by the balance between mutation and selection, the selection has to be different between common and rare disease. An explanation would be that the selection is intense towards rare disease because it would be reproductive lethal whereas selection might be mild towards common disease that only occurs later on.

Now, if we consider that $f = f_0 = f_{exp}$, in the ancestral population, all disease loci had a simple spectrum. The number of disease alleles should be $n = 1.1$ for both the rare and common disease since $1 - f_0$ is close to 1. A single disease allele accounts for 90% of the disease class. In a modern population size, all disease loci should have a complex spectrum. The difference between the common disease and the rare one is the speed at which they reach a complex spectrum.

Each generation, $\frac{(1-f_0)\mu}{f_0}$ of the alleles in the disease class are expected to be

replaced. The proportion of original alleles that are expected to remain after t generations is $e^{-(1-f_0)\mu t/f_0}$. The half-life of allelic replacement is $ln(2)f_0/\mu(1-f_0)$ generations. Depending on the value of f_0, this half-life varies greatly from thousands of years for the rare disease to million of years for a common disease.

This model relies on strong assumptions as one can see. First of all, the human population did not grow from few thousands to a billion instantaneously. With a gradual population growth, the growth in diversity would also be slower. Lander claims that it does not change the final results very much but it remains an approximation. The ancestral population is also assumed to be of constant size to allow ϕ to be at equilibrium but the ancestral population might very well have fluctuated in size and have had an influence on allelic diversity. There are neither structure in the population nor selection pressures. Those factors may greatly influence the frequency of alleles in a population.

2.2.2 Debate

The common disease common variant hypothesis (CDCV) is probably one of the most debated inside the population genetics community. One argument for CDCV is that stochastic phenomena or purifying selection would get rid of rare, disease-causing variants. The opposing argument states that there is a large population with common diseases, so less susceptible to stochastic phenomena that knock out rare variants. Also, there has been a recent explosion of the human population that may come from a bottleneck. Furthermore, selective pressures against many modern diseases that are associated with abundant access to food

and a sedentary lifestyle, have only been acting for six to eight generations. While we are learning about common disease loci and variants from association and linkage studies, the genetics of common traits is likely more complex; relatively rare alleles with relatively weak effects probably also play a role.

There are other explanations besides CDCV. One is allele heterogeneity, where there are multiple alleles at a single locus that are weak individually, but which, when aggregated, have a frequency high enough to explain a disease. Another explanation is locus heterogeneity, where there are many loci that confer risk, and any individual with a disease will have a small fraction of the risky loci. Also, carrier status of Mendelian variants may contribute to common, complex diseases. Studies can look at a mendelian gene for study on related complex trait. Mutation selection is weak selection against predisposing variants; so it is possible that the presence of common diseases is due to new mutations. In some cases, environmental factors leading to disease were not common until the past few hundred years. Arguments for or against CDCV based on natural selection are dangerous. Selection may be acting for a different, non-disease related aspect of the variant, conditions in the past may not have allowed many people to develop disease associated with the variant, and a potential founder effect may have preserved genes with moderate effects, even if they are common. Another important fact to remember when studying complex traits is that even strong alleles can be affected, and their effects can be entirely reversed in some cases by other genes at other loci or by the environment.

Environment affects common diseases more than some common variants. Some

authors argue that CDCV doesnt work with the multi-regional theory of human origins, because CDCV depends on the idea that this common variant arose long ago (hence its ability to become common), got fixed in the population, and any other competing alleles that could also cause disease at that locus have not had enough time to significantly alter the predominance of the common allele.

Current knowledge is based on the studies that have been performed, and these studies may be biased towards finding common variants, due to limitations in sample size, study design, or technology used. These constraints lead to the challenge of how we can detect rare allele effects. Whole genome sequencing may be a good technology to use for detecting rare alleles, but as of now it is too expensive and time consuming to carry out studies on a large scale. The question of CDCV versus genetic heterogeneity needs much more work. One thing to remember is that mutation rates vary among loci, and if a locus has a very high mutation rate, there may be enough heterogeneity to decrease the utility of linkage mapping.

Even if a common variant is involved with a common disease, other, rarer variants may also play a role. In cases where the common variant is known, it is important to remember that the variant can have effects on systems not directly involved in the disease in question, so it may have an impact on other diseases as well. Neither the CDCV nor the CDRV (common disease rare variant) theory has anywhere near enough evidence behind it to support its predominance. It is hard to decide when a gene or variant can be nominated as CDCV or CDRV, because many studies have not gone far enough beyond linkage associations or do not

have large enough sample sizes to settle the question for a given variant. There may also be rare variants with strong effects producing a different variant of a common disease. These variants would be easily missed with association studies. Lack of standardization of approaches to linkage or genome-wide studies makes it difficult to compare results in multiple studies.

CDCV is certainly true for some versions of certain diseases, but the heritability of common diseases is almost certainly some mixture of common and rare variants; and the only question is what is the best way to model this mixture. Common diseases tend to have a genetic component, but they tend to be polygenic and interact extensively with environment, making it harder to understand the nature of that heritability. Also these diseases may represent more than one physiological pathway. One thing to beware is that a gene can have a strong effect on disease development within an individual or even within a population, but that it may not add much to a familys risk, due to various factors including other genes that are present in a familys genome. For this reason, if heritability is defined as the risk of a disease based on family history; it may mask heritable effects of certain variants.

Rather than looking at the relative frequency of a locus and deciding whether that locus constitutes a common variant or not, it has been posited that the null hypothesis when examining the question of CDCV should be that the allele spectrum of a disease gene resembles the average allele spectrum of the human genome. This assumption is based on knowing the overall human genome allelic spectrum.

2.3 Simulations

The model we have implemented is still rudimentary and we cannot yet try realistic scenarios. Nevertheless, we can see if known features of population models are respected and what some disease models have to show us. First we can see what happens with a population of constant size. We know that we should reach the Hardy Weinberg equilibrium after the first generation. Since at the first generation, everybody is the same (and has no mutations), we run the simulation over 1000 generations, of which the first 200 generations have mutations to create a diverse population. After the 200th generation, there are no more mutations added. We follow two types of mutations. The blue curve represents a SNP that is selectively neutral but that will be responsible for a common disease later on in the life of the individual. The green curve represents the number of people who have a heterozygote advantage. In the simulation, being a heterozygote gives you a chance of 1 to be kept if selected, while if you are homozygote, you have a chance of 0.8 for the wild case and a chance of 0.01 for the mutant case.

What we see in Figure 2.1 is that the number of people with the heterozygote mutation grows up to a certain threshold and then remains constant (the small variations are due to genetic drift). This can be explained by the fact that if there are too many heterozygous individuals, the population reaches a point where a lot of the mutant homozygotes are produced. These homozygotes will most likely die and therefore diminish the presence of the mutant allele. This acts as a control mechanism. The blue curve is a little bit more unusual. As said earlier, after the 200th generation, it should reach an equilibrium. The reason it does not is

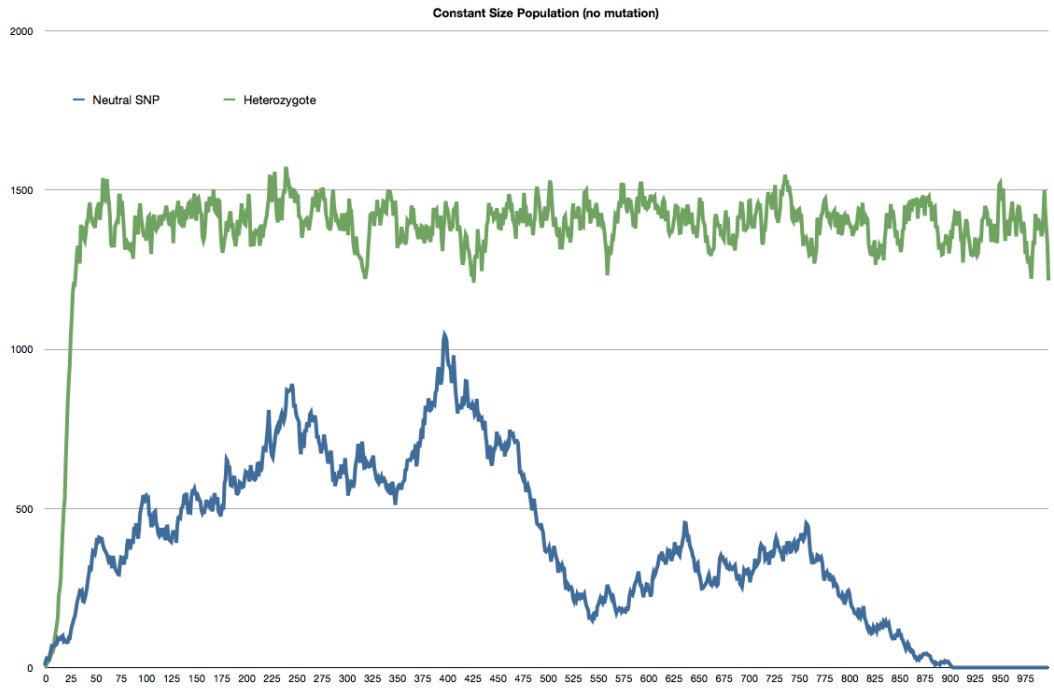

Figure 2.1: Constant size population of 5000 individuals with no mutations after the 200th generation

genetic drift. If the population is small, the Hardy-Weinberg equilibrium may be violated. Genetic drift can eliminate certain members out of proportion to their numbers in the population. If this is the case, an allele might start to drift toward either fixation or extinction. Here, the simulation clearly shows an extinction of the SNP.

Now, if we had not stopped the mutations from occuring, the results would look like the Figure 2.2. In this case, mutations keep occurring at a Poisson rate so the blue curve, which is not under any kind of selection other than genetic drift, has a linear progression with fluctuations due to genetic drift. What is interesting is that the green curve follows the same pattern as it did without mutation, reach-

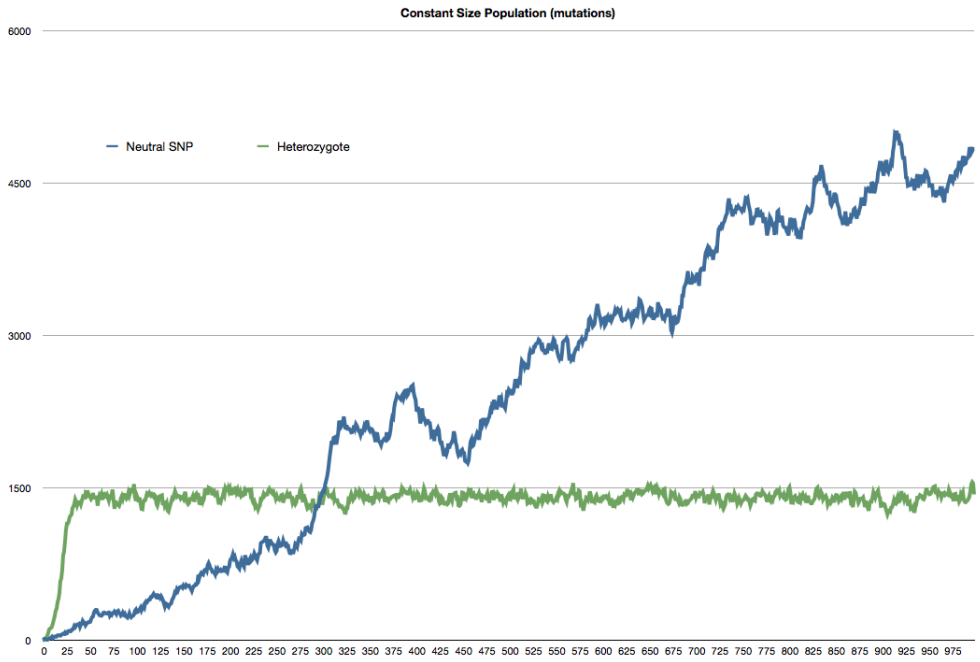

Figure 2.2: Constant size population of 5000 individuals with mutations

ing an equilibrium of around 1500 heterozygous individuals out of 5000. This threshold depends on the parameter. The more being heterozygote is selectively positive, the higher the number of heterozygote individuals will be and conversely the less advantage conferred by heterozygosity, the lower the proportion of heterozygotes in the population.

We have been following a single SNP so far. But we know that there is probably a mixture of different SNPs that are responsible for most common diseases. In the Figure 2.3, we ca see the behavior of the blue curve. It still represents the number of SNPs present in the population, but this time, we follow ten SNPs. Here, as long as an individual has fewer than four of those SNPs, he is fine but if

he has more than four, this individual cannot be picked up in the next generation. It is worth noting that, like the number of heterozygote individuals, the number

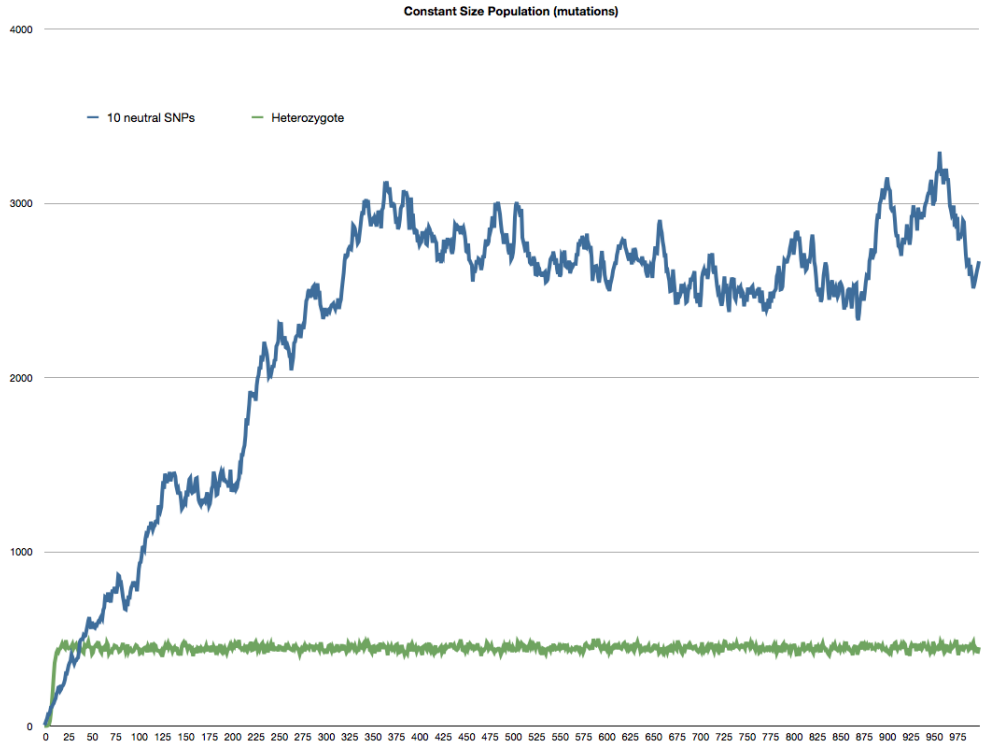

Figure 2.3: Constant size population of 1000 individuals with mutations, following ten SNPs

of SNPs reaches an equilibrium at a certain point.

In Appendix C, results with different population size dynamics are displayed. As stated earlier, there are still many features that can be implemented in the model in an easy way. Adding population structure with some type of island model is a possibility. It will also be interesting to define a weight for each SNP that we track in order to refine the cumulative effect of their presence in a genome instead of just a strict threshold. Another extremely important feature is to track LD.

Population	Generations	u	r	Genome Size	Time
100	1000	0.1	0.1	100 000	0m33s
100	1000	0.1	0.5	100 000	0m26s
100	1000	0.5	0.1	100 000	2m00s
100	1000	0.5	0.5	100 000	2m09s
1000	1000	0.1	0.1	100 000	5m02s
1000	1000	0.1	0.5	100 000	4m36s
1000	1000	0.5	0.1	100 000	21m54s
1000	1000	0.5	0.5	100 000	22m03s
1000	1000	0.1	0.1	1 000 000	5m59s
1000	1000	0.1	0.5	1 000 000	4m19s
1000	1000	0.5	0.1	1 000 000	23m36s
1000	1000	0.5	0.5	1 000 000	25m58s
1000	100	0.1	0.1	1 000 000	0m7s
1000	100	0.1	0.5	1 000 000	0m8s
1000	100	0.5	0.1	1 000 000	0m17s
1000	100	0.5	0.5	1 000 000	0m19s

Table 2.1: Running time of simulations with different parameters. u is the mutation rate per generation per sequence and r is the recombination rate per generation per sequence. The simulations were run on a 3.06 GHz Intel Core 2 Duo with 4 GB of RAM. The code is written in python and interpreted using pypy.

For now, we have hot and cold spots for recombination, which is the first step, but we need a tracking implementation to see how LD plays a role.

In terms of performance, as discussed previously, using a Wright-Fisher approach is not the fastest way to solve the problem. Nevertheless, the running time for the simulations is more than acceptable. Different running times for different parameters are displayed in table 2.1. We can see that the mutation rate is an important factor in terms of running time as well as the population size while the recombination rate and the size of the genome do not seem to play a big role.

Chapter 3

Genome Wide Association Study

3.1 Status of GWAS

There are two main approaches to connecting the genes involved in common diseases. These include 1) the candidate gene study in which one can use either association or re-sequencing approaches, and 2) the genome-wide study in which one uses linkage mapping and the genome-wide association (GWA) study.

Until recently, genome-wide linkage analysis was the main method used to identify disease genes. It has been successful for mendelian diseases (where only one gene is involved) [**?**] where there is near a one to one connection between genotypes at a single locus and the observed phenotype. The most famous successes are cystic fibrosis [**?**], Huntington's disease [**?**] or Duchenne's syndrome [**?**]. Those studies have also had some positive results for common diseases in cases such as schizophrenia [**?**], Crohn's disease [**?**] and type 1 diabetes [**?**] , but for most common diseases, the results are far from being successful [**?**]. Many factors

can explain this lack of predictive power. Most complex traits have low heritability, phenotypes of those diseases are hard to define precisely [?] and finally, the design of the study itself [?] is often flawed. It is argued that with bigger samples [?], larger pedigrees [?] or dense marker sets [?, ?] linkage analysis could give better results. However, candidate gene studies are still required to move from a wide region of linkage to the causal gene(s) within this region. The biggest problem lies elsewhere. Linkage analysis cannot efficiently identify common variants that have moderate effects on disease [?, ?]. For most common diseases, their phenotype is composed of a combination of multiple genetic and environmental factors and their interactions [?]. Each individual variant will account for a small part of the phenotype of the disease. Whether the CDCV hypothesis is true or rare alleles also contribute to common disease, the poor power of linkage analysis to detect alleles with low penetrance make them unsuitable to use them alone for finding alleles that are susceptible to take part in a disease.

A candidate gene is a gene for which we have evidence or at least a strong indication that it plays a role in the trait or the disease that is studied. One type of analysis of candidate genes is done by re-sequencing the entire gene in the studied populations (often case and control) and looking for variant(s) between the populations. The main problem with this approach is its cost, effectively limiting the regions where to look for the candidates (usually in the coding regions). We can also use association studies with candidate genes. They are cheaper and simpler than their resequencing counterpart and have been proposed to find common variants that underly complex traits. Basically, an association study compares

the frequency of alleles of a variant between case and control. Candidate gene association studies have identified many genes that are partially responsible for common diseases [**?**, **?**, **?**]. Still, candidate gene studies require to have some biological evidence implicating it in the disease trait. Even if hypotheses made on those genes may be very broad (for example, that a gene is somehow involved in a certain pathway), it is impossible to overcome the fact that only a small fraction of the genetic risk factors will be determined. Worse, this approach is clearly inadequate in the case that the physiological defects of a disease are unknown, therefor no assumption can be made .

A GWA study is defined by the National Institute of Health as a study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits. A GWA study can be decomposed into four parts. First, the selection of a large number of individuals for both the case and the control group. Second, the genotyping quality must be high, implying the use of DNA isolation, genotyping and data review. Third, statistical tests have to be run for association between the SNPs passing a certain quality threshold and the disease. Finally, the experiment should be replicable in an independent population sample. Even if the primary goal of GWAS is to detect SNPs associated with a disease, this technique also permits identification of variants relative to quantitative traits such as height [**?**]. It can also demonstrate gene-gene interactions (as with $GAB2$ and $APOE$ in Alzheimer disease [**?**]). It can also detect high-risk haplotypes inside a single gene (as in atrial fibrillation [**?**]).

As stated earlier, the design of a GWAS often includes two populations; a case

population which is formed of theindividuals affected by the studied disease and a control population with healthy people (who are not affected by the disease in question). Allele frequencies between those two groups are then compared. This design is the simplest but also the one with the most assumptions. As usual, the more assumptions that are made, the more bias is introduced [**?**]. Another study design is called the trio design. In a trio design study, the parents of the affected patients are included in the population. Only the offspring needs to display the phenotype of the disease but the three individuals will be genotyped. Also, the disease variant(s) is transmitted in excess of 50% to affected offspring from heterozygous parents. A last design is the cohort design. It implies an extensive collection of baseline information about the studied population. Those individuals are then observed prospectively to assess the incidence of disease in subgroups defined by the variants. Each of these designs has advantages and drawbacks. For the case-control design, advantages include simple implementation. It yields results faster than the other designs. It is also easy to gather large population for the groups and in term of epidemiology, this design is optimal for studying rare diseases. On the other hand, this design is prone to biases such as population stratification. The case group is often made of prevalent cases which does not take into account the variety of disease expression (like fatal, short, mild or silent cases). It also tends to overestimate the risk for common diseases. A major advantage of the trio design is resilience to population stratification since the population structure is controlled. In addition, during the genotyping quality control phase of the study, we can check for Mendelian inheritance patterns and

trio studies do not require phenotyping the parents. The trio design is useful to examine the children's conditions. But it is hard to unite parents and children with late onset diseases and this design is extremely sensitive to genotyping errors, imposing higher standards for quality checks. The cohort design, unlike trio studies or case-control studies, permits direct assessment of disease risk. Since cases are developing during the observation, they are free of survival bias even if some other biases can still exist (though to a lesser degree than in the control-case design). Unfortunately the logistics of cohort studies pose some difficulties. One needs a large sample for genotyping if the incidence is low. Cohort studies are notoriously expensive and require a long time for observation. It is not always agreed upon whether the consent obtained during the study is sufficient for data sharing. Cohort studies also need variation in the studied phenotype. In contrast to the case-control design, it is very poorly suited for studying rare diseases.

The first step in a GWAS is to chose a case and a control group. The difficulty in choosing subject to place in these groups lies in the misclassification of individuals inside the case group (healthy people put into this group). Such misclassifications lead to a loss of power. Misclassification, however, is difficult to avoid, as the genetic architecture of complex diseases is poorly understood and accurately diagnosing those diseases can be difficult making the marking of individuals a complex process. For the control group, the individuals should de taken from the same population as for the case group and should also have the possibility to develop the disease. For example, putting a woman in the control group of a disease that only affects men would be problematic since she cannot develop the

disease. In some cases, she may have the disease trait but is lacking the necessary conditions to trigger the disease, as those conditions may be coded on the Y chromosome. In this situation, the control group is mixed with latent cases. For the study of common diseases such as coronary heart disease, the control group must truly be free of disease. Still, the Wellcome Trust Case Control Consortium seems to lean in the direction that the"quality" of the control group does not interfere much with the discovery of variants associated with the disease. There is also a consensus that the larger the sample size in a case control study, the better the results will be. The population stratification (or structure) can also be resolved by different techniques if the case and control groups are well matched for wide ethnic background. Still, those techniques do not get rid of all the biases introduced by population stratification.

The second step is to control the quality of genotyping. GWAS rely on a strong linkage disequilibrium among SNPs. Genotyping is performed either on chips or arrays and the genomic coverage of those platforms is often assessed by the percent of common SNPs having an r^2 value (as defined in chapter 1) of 0.8 or bigger. Depending on the population, the number of SNPs that are tested on those genotyping platforms will represent a greater or smaller proportion of the common SNPs variations in that population. For platforms with 500k to a million SNPs, 67 to 89% variations can be captured for European and Asian populations while only 46 to 66% for the African one [**?**]. It it possible to use higher density platforms. Recently, on top of the SNPs, those high density platforms have added probes for copy number variants (CNV) which have become of great

interest because of their apparent ubiquity and potential dosage effect on gene expression [**?**]. Still, while capturing SNPs and CNVs, there are still features like inversions, insertions and deletions that are hard to capture. There are no universal quality-control thresholds to define a set of good genotypes. Depending on the focus (accuracy or call rate) of the study, the threshold will be different. If you want high accuracy, the threshold for calling genotypes will be high and therefore many SNPs will have a low call rate, leading the researcher to discard some of the true signals. On the contrary, if the focus is on call rate, the study will end up with a number of poorly performing SNPs that will resist the phase of quality-control. The remaining samples undergo other checks to filter genotyping errors. If SNPs are significantly in violation of the Hardy-Weinberg equilibrium, they can be discarded. For the trio design, the mendelian inheritance errors are checked.

The third step is the statistical analysis of GWAS. There are some tools that allow representation of the data from GWAS, one of the most common being the quantile-quantile plot. On those plots, we can see if the study has had results that are more significant than results expected by luck. The most used and arguably powerful tool to analyze results of GWAS is a single-point, one degree test of association, such as the Cochran-Armitage test. Basically, the genotypes of case and control groups are compared SNP by SNP with or without adjustment for relevant covariates (like the principal component of population substructure). It is robust to small variations from additivity on the logistic scale. The use of alternative models such as general, dominant or recessive might increase the detection

of some signals but the calculation of type 1 error rates might get complicated with multiple correlated tests. The most widely used model is an additive one where each copy of the allele accounts for the same increased risk of disease. We can compute odds ratios of disease associated with the risk genotype(s). It is also possible to compute risk due to membership in a specific population. The problem of those values is that they are often overestimated because odds ratios increase relative risks needed for population attributable risk calculations. This initial overestimation of odds ratio tends to create problems when trying to replicate a study because larger samples are then needed to detect smaller odds ratios. To assess the significance of genotype association findings, the classical statistical approach based on p value prevails. The problem is for classical values of p (such as ≤ 0.05) of significance, the number of SNPs associated with a disease will be extremely large (in the order of 10^5). Obviously, almost all of those SNPs are false positives. To deal with this problem, people often use the Bonferroni correction (the p value is divided by the total number of tests) to decrease the rate of false positives. This correction, while commonly used, is undermined by the fact that it assumes an independent association of each SNP with the disease while it is known that SNPs are correlated through LD. Those limitations have lead to the development of other techniques, mostly based on a Bayesian approaches, with an integration of the likely number of true positives and the power of a given study [**?**, **?**]. To improve the power of a study, we can also use haplotype based and imputation methods [**?**, **?**]. The improvement comes from the fact that the coverage of common variants provided by the GWA platforms is not complete.

The last step is the replication and validation of the study. Because of the high number of false positives, an effective way to test for real associations is to replicate the results with independent samples [**?**]. This analysis could be done in a single GWAS with a multistage design or could be reported separately. To replicate studies, one accepted method is to study the closest possible phenotype and population to the original study and demonstrate a similar magnitude of effect and significance for the same SNP as the initial report [**?**]. Some relaxation of those conditions can be tolerated such as use of different populations (European then European plus African) or related phenotypes (such as fat mass in addition to obesity), or different study designs. It is common for a study not to be reproducible. Many factors can explain this, such as population structure, selection biases, phenotype definition differences, genotyping errors, etc. One way to solve these differences with the original study might be to use larger samples although it is not always possible.

3.2 HapMap

After reviewing the state of GWAS, it is quite clear that something else is needed if we want to be able to find variations that are related to disease. Linkage studies are extremely powerful when it comes to Mendelian diseases but are inefficient when the the effects of different variants on a disease is diluted among all of them. It is hoped that association studies will overcome those problems but no real breakthrough has been seen yet. The single point analysis presents too many

flaws to be of a great help. This is why people have started to lean toward more complex analyses, taking into account not one SNP but many. This set of SNPs is known as haplotype. Before going into more details about what are haplotypes and how they could help detecting variants linked to complex diseases, let us introduce a project that aims to help with the use of haplotypes in GWAS.

The International HapMap Project is composed of a consortium of scientists from different countries. The project is based on the premise that 90% of human genetic variation is due to common variants of about 10 million SNPs [**?**, **?**]. In addition, most variants have individually arisen from a single historical mutation rather than being the products of multiple independent mutations, due to the low mutation rate at a given site in the human genome.

Over time, as SNPs accumulate, each new SNP would be associated with SNPs that arose prior to it, leading to linkage disequilibrium between a certain allele of one SNP and alleles of neighboring SNPs. Governed by the nature of linkage disequilibrium and recombination events, the farther apart two SNPs are, the less likely they are to be reliably associated due to LD. The sequence of neighboring SNPs constitute a haplotype, and because of the linkage between SNPs, the HapMap project constructed haplotypes and identified tag SNPs i.e. identifies a few SNPs out of the many in a region of a chromosome that are common and therefore older than other SNPs. Based upon the sequence of the tag SNPs, the project predicts the nearby SNPs by comparing the tag SNPs to a haplotype map. The project estimates that 200000 to 1000000 SNPs will suffice to predict the sequence of all 10 million common SNPs in an individual's genome.

The purpose of the HapMap project is to identify areas of common variants on the human genome, and to create a database of these variants, as well as identifying suitable tag SNPs and suitable other SNPs that have a high degree of linkage disequilibrium with the tag SNPs. Both their locations and sequences will be useful for future studies examining the association between diseases and certain haplotypes. To that end, data is planned to be made completely available in a timely fashion for other researchers to use. The study gathered data from populations in Utah (of northern and western European descent), Ibidan Nigeria, Beijing and Tokyo. Despite the selection of various populations, most haplotypes were expected to be found in every population.

The project aims to genotype 600000 evenly spaced SNPs in an initial round of genotyping, each SNP with an allele frequency $\geq 5\%$, with priority given to SNPs that would change amino acid sequence in a gene product, SNPs that have been validated in previous studies, and SNPs that are found independently in two or more samples. Associations of LD between these alleles will be analyzed. Further sequencing will identify other, less common SNPs in areas of poor LD.

3.3 Haplotype: The Missing Link?

As we have discussed, as haplotype maps became available, and researchers were no longer limited by the analysis of single SNPs, there was hope that GWAS would finally allow us to discover the secret behind complex diseases. Yet, the HapMap project and the GWAS that followed didn't bear the fruits that were expected. The question now is to assess if those fruits are not ripened yet or if

they are just not what we were expecting them to be. Here, we will focus on the problems that crop up when using haplotypes.

Since we are talking about complex diseases, usually more than two loci are studied together. In this case, we try to distinguish between pairs that have high levels of LD from those that do not [?]. The results are often displayed as a graph to describe patterns of LD in the genome. Those highly correlated SNPs form groups that are usually referred as haplotype blocks. It has been noticed that the boundaries of these blocks were correlated with hot spots of recombination. Inside a block, the recombination rate is low while it is much higher in between the blocks. It is now hypothesized that the human genome has a block-like pattern of LD. The size of those blocks varies from few kb to 100 kb [?]. The view of the genome as partitioned into haplotype blocks is recent. Before that, the most common belief was that, under assumptions that tried to fit the history of modern human evolution, the further apart SNPs were on chromosome,the less LD they had and little LD would be expected for SNPs distant by more than 3 kb [?]. The structure of genomes into haplotype blocks has changed this view and it is now believed that LD is effective over much longer genome distances (to the order of 10 or 100 kbp). It is also hypothesized, and applied in the HapMap project, that the study of only one SNP inside a block might be sufficient to reveal association with all other SNPs within the block. This would allow significant reduction in the number of needed SNPs to perform association studies, therefore making it more affordable [?]. The reality is less idyllic because some regions of the genome cannot be described with this block structure. There is also not

a single way to define haplotype blocks, changing the boundaries of those blocks hence changing associations between those blocks.

The major setback when studying haplotypes is called the problem of unobserved haplotype phasing. On a theoretical level, a value such as D assumes that the haplotype of an individual is available. In reality, only diploid genotypes can de found. Let us imagine surveying three loci in three individuals who are going to be genotyped. If the genotype of the first individual is $AaBBcc$ then his haplotype is obvious and there is no problem determining it. His haplotype is ABc and aBc. As long as only one of the loci is heterozygous, there is only one solution to resolve the haplotype without uncertainty. Now, individual 2 has $aaBbCc$ for a genotype. To determine his haplotype, more information is needed. Indeed, just with this genotype, this person could have the following haplotype aBC and abc or another one aBc and abC. The number of possible haplotypes for a person increases exponentially with the number of heterozygous loci that are studied. In our example, if a third individual had three such loci (the genotype is $AaBbCc$), he would have four possible haplotypes: ABC and abc, ABc and abC, AbC and aBc, or aBC and Abc. There is a need of methods to determine the correct haplotype from genotype data. This problem is called resolving haplotype phase. One of those methods involves genotyping the parents along with the individual of interest. Going back to our example, if the genotypes of the parents of the second individual are $AaBBCc$ and $AaBbcc$ then person two has to have aBC and abc as haplotype. On the other hand, if the parent's genotypes are $AaBbCc$ and $AaBbcc$, we still cannot resolve the haplotype phase. More commonly, sta-

tistical imputation methods are used to infer haplotype phase and then inference is used as data. There are numerous methods that have been developed based on different concepts such as maximum likelihood [?], parsimony [?], combinatorial theory [?] and a priori distribution derived from coalescent theory [?]. The main idea behind these theories is that people who have at most one heterozygous locus among all the studied loci provide some information about haplotype frequencies. This information is then used to infer the haplotype phase of the other individuals. This approach has been reasonably fruitful in term of results, especially for common haplotypes. Still, it ignores the uncertainty that defines the inference step. Inferred frequencies of rare haplotypes can be quite inaccurate [?].

The discovery of some block-like structure within the genome has shown that regions that are far apart can still be in LD and are therefore important to understand. The hopes that rose with the study of haplotypes have been shattered due to a simple fact: with current techniques it is impossible to resolve the haplotype phase with certainty. As long as this issue persists, there is little hope that haplotype analysis will be useful in association studies. There is one way to resolve the haplotype phase with certainty: directly sequencing the haplotype. Unfortunately, as of today, no sequencing technology allows haplotype sequencing. We are now going to review existing different technologies and propose a novel scheme that will permit us to sequence haplotypes and therefore might be a major breakthrough in sequencing technologies as well as in population genetics.

Chapter 4

Sequencing Technologies

4.1 Technologies

4.1.1 Sequencing

Sanger: Capillary gel electrophoresis

Sanger sequencing was developed in the 1970s at the same time as Alan Maxam and Walter Gilbert devised a different sequencing method. In the modern version of Sanger sequencing, cloned DNA (originally cloned using bacteria, but now usually amplified using PCR) is primed and dideoxribunucleotide triphosphates (ddNTP) are added to the reaction mixture (A,C,T, or G), along with normal deoxynucleotides of all four bases. The ddNTPs are labeled using a fluorescent dye, with a different color used for each base. Using a DNA polymerase, a base is added to each cloned strand until a ddNTP is incorporated, and the resulting strands are run through a sensitive electrophoresis gel, capable of resolving dif-

ferences of one nucleotide between strands. For every given length strand, the fluorescent label is detected, and based upon the color of the label, the base at that position is recorded.

Sequencing by synthesis

The Sanger method is based on chain termination and separation in capillary gel. In sequencing by synthesis, cycles of the four nucleotides are consecutively added, a nucleotide is incorporated, it is detected, and the chain is continued, such that there is no need to use the electrophoresis step. In addition to pyrosequencing, sequencing by synthesis is used commercially in an array format, where fragments are produced, amplified, and hybridized to an oligonucleotide that is linked to a glass surface. The strands are denatured, primed and 3- blocked fluorescent-labeled deoxyribonucleotides are added sequentially. After each addition, the surface is washed to remove unincorporated nucleotides and any incorporation is detected, followed by deblocking the 3- end, and adding the next nucleotide.

Sequencing by ligation

DNA ligase is an enzyme that links together double-stranded DNA or can even link together one of two strands of DNA. The enzyme is quite specific and will not link together mismatched strands, a feature which is helpful in preventing formation of malformed or mutated DNA during reproduction. This method utilizes the fact that DNA ligase, the enzyme that can link double strands of DNA, or even one of two strands of DNA is highly specific and tends not to link together mismatched bases. In polony sequencing, a query fragment is amplified

and hybridized to an anchor primer. A group of random 9-mers is then added, with a fluorescent label at a specific base position. As with modern Sanger sequencing, each base has its own color. A detector then reads to see which color predominates at the given base position, and the complex is stripped apart and 9-mers washed away to reset for the next cycle, which will look at the next base position.

Sequencing by expansion

This technology converts DNA into an Xpandomer, which encodes sequence information with low noise, allowing for reduced sample preparation and processing time. In May of 2011, Stratos Genomics received a patent for a method of converting DNA to an Xpandomer.

Sequencing by hybridization

The principle of sequencing by hybridization rests on the fact that complementary single strands of DNA will hybridize if put in proximity together. If oligonucelotides of known sequence are mixed with fragments of unknown sequence, one can determine the sequence of the unknown strand by determining which oligonucleotide has bound the unknown fragment. Currently, this type of sequencing is used to test for SNPs, by having arrays of similar oligonucleotides, and adding fragments from a specific site in the genome/chromosome [**?**, **?**, **?**].

Pyrosequencing

Pyrosequencing is basically a modification of sequencing by synthesis in which a primer is hybridized to an amplified template and mixed with DNA polymerase, ATP sulfurylase, luciferase, and apyrase. Each of the four dNTPs is added individually, in a cycle, and when an NTP is incorporated, the ATP sulfurylase converts the released inorganic pyrophosphate to ATP. The ATP then allows luciferase (an enzyme present in fireflies) to convert luciferin to oxyluciferin, which produces visible light. The apyrase serves to reduce the amount of false signals that can be caused by natural dATP. The amount of inorganic phosphate released, and therefore, the amount of visible light produced, is proportional to the number of nucleotides incorporated. In other words, if four of a certain base are incorporated in a row, the signal will be higher than that for three or fewer. The light is detected by some sort of photon-detection device, and is displayed as a peak on a pyrogram, or flowgram.

Ion semiconductor sequencing

This is another technology that is derived from sequencing by synthesis during which a complementary strand is built. This technology is based on a well-known biological fact: when a nucleotide is added into a strand of DNA by a polymerase, a hydrogen ion (H+) is liberated. Ion semiconductor sequencing will basically detect the release of this hydrogen ion. A semiconductor chip is made of a high-density array of micro wells. Each of those wells is filled with a single-stranded template DNA and a DNA polymerase. Then, those wells are flooded with A, C

T and G dNTP sequentially. Under the wells, there is an ion sensitive layer and beneath this layer there is an ion sensor. If a C is added to a DNA template and is then incorporated into a strand of DNA, an ion will be released. The charge of this ion will change the pH of the solution and the hypersensitive ion sensor will detect this variation. Each nucleotide addition is directly recorded, without the need of scanning or camera or light.

Nanopore sequencing

When a channel has an electrical voltage applied across it, and there is a particle pulled through that channel, the current will decrease. This is the basis of nanopore sequencing in which DNA is drawn through a channel that is protein-based or synthesized. The benefit of nanopore technology is the potential for long read lengths and the possibility to cut out the DNA labeling step. It would allow very high throughput due to the small size of the nanopores, at a relatively low cost. So far, it has proven difficult to distinguish individual nucleotides as well as to force DNA through the channel without the molecule folding into its characteristic hairpins and loops.

4.1.2 Mapping

Optical Mapping

This single molecule technology is based on a de novo process that generates a high-resolution, whole genome and ordered restriction map. It works with a single molecule, is independent of sequence information and does not require

amplification or PCR steps. The idea is to map the location of restriction enzyme sites giving the output a resemblance to a bar code (a black line appears where a restriction site is found). There are five steps in order to get an optical map. The first step is to extract the DNA from the cell. Once this is done, single molecules of DNA are stretched and immobilized on a surface. The DNA can be held by electrostatic interactions on a positively charged surface or along microfluidics channels. The next stage is to digest the molecule with restriction enzymes. Those enzymes will cut the molecule at their digestion sites. The resulting fragments remain attached to the surface so they keep their order. Since the DNA has some elasticity property, it shrinks back a little at the ends of those sites, leaving a gap between fragments which can be detected with optical microscopes. After the digestion is done, the DNA is stained with a fluorescent dye. In order to determine the size of a fragment, the intensity of the fluorescence of each fragment is computed. At the end of this process, we have a single molecule map. Finally, all individual molecule maps are assembled by overlapping fragment patterns to obtain a consensus, genomic optical map.

BioNanoGenomics

4.2 Assemblers

4.2.1 Phrap

There is no publication about the algorithm behind Phrap even though it is one of the most widely used assemblers. We have to go to the website http://www.phrap.org

to find a description of the algorithm. It is decomposed into five steps. First, a sorted list of fragments of at least a minimum length is created. Second, for each pair of fragments, a band around a diagonal that is defined by matching fragments is defined and overlapping fragments are merged. Phrap uses an implementation of the Smith-Waterman algorithm called SWAT to identify matching segments above a certain score. SWAT is recursively applied between matches by masking out the current matched regions. Third, two hypotheses are tested and compared through a log-likelihood ratio. The first hypothesis is that the reads truly overlap and the other hypothesis is that they are from repeats of 95% similarity. A positive log-likelihood confirms the first hypothesis while a negative one confirms the second hypothesis. Fourth, a fragment layout is progressively generated using a sorted list of matches in term of their log-likelihood scores. Finally, a consensus sequence for each contig is built using a a weighted graph (using a single source maximum weight path algorithm) with selected positions of matches as vertices.

4.2.2 TIGR

The first bacterial genome, H. influenzae, was assembled by TIGR [**?**] using the shotgun strategy in 1995. This assembler follows two phases, first a pairwise comparison of the fragments and then an assembly of those fragments. After the pairwise overlaps between fragments have been computed, a fragment is merged with the current assembly if it satisfies four conditions. The overlap has to be bigger than the minimum overlap length defined, there has to be more than a minimum similarity in the overlap region (defined as a percentage of the best

possible score), the length of overhang (the region in the alignment where two fragments do not match) should not exceed a certain maximum and there should be no more than a certain maximum of local errors. The maximum error threshold is used to discard overlap with clustered errors but have passed the similarity test.

If a fragment passes all those tests, it is added to the current assembly. No consensus is computed then but TIGR keeps a trace of what bases have been aligned to that position. It keeps a record of bases and gaps in a profile for each position. After the assembly is done, a consensus sequence is generated using this profile, choosing the most frequent bases. Fragments that have a number of potential overlaps based on pairwise comparisons are labelled as repeats. When such a fragment is incorporated to the assembly, the match criteria is increased (the similarity test) to distinguish inexact repeats. Since it is still impossible to avoid false overlap when repeats are longer than the fragment size, TIGR incorporates mate-pair information as well to deal with repeats.

4.2.3 CAP3

CAP3 is the latest version of the CAP [**?**] assembler. In CAP2 [**?**], some improvements had been developed such as filtering potentially non-overlapping fragments, identification of chimeric fragments (using an error rate vector for each fragment) and handling repeats by constructing repetitive contigs while merging two different contigs. In the third version of the software, other improvements have been created. Now, 5' and 3' poor quality regions are clipped. It is done by

using both base-quality values and sequence similarities. A good region of a fragment is defined as one with any region of at least a minimum size of high quality values and any sufficiently long region that is highly similar to a high-quality region of another fragment that can be defined as good. The 3' and 5' clipping positions of a fragment are determined by the boundaries of good regions.

The alignment between two fragments is determined over a band defined by the optimal local alignment while clipping the poor quality regions. Then the quality of the overlap is assessed by five different measures: minimum percent identity, minimum length, minimum similarity score, difference between overlapped fragments at high-quality bases and difference between the expected sequencing error rate and the error rate of the treated fragment. While contigs are built, CAP3 uses mate-pair constraints. An initial layout is built greedily in decreasing score of overlaps. Then this layout is tested by mate-pair constraints. The region with the largest amount of unsatisfied constraints is located and those constraints are checked for being satisfiable by aligning unaligned pairs according to their distances. If this is possible, corrections to the region are made by adding satisfiable pairs and breaking unsatisfiable ones. The new layout is then retested until such regions cannot be found and the program stops. Finally, contigs are ordered and linked with unsatisfied constraints (for example, using mate-pairs in two different contigs).

4.2.4 Celera

Celera was the first assembler to successfully assemble reads from large eukaryotic genomes ($\geq$ 100Mbp). It not only uses mate-pair information to resolve the repeats problem but also uses available external data in order to get the best possible assembly of the genome. This assembler has a different level of "aggressiveness" to treat the reads, starting from the safest moves and progressing to bolder ones. The Celera assembly is divided in five steps. The first step is called screener and essentially serves to treat repeats. Each input fragment is checked for matches to known repeat regions and is either marked (soft screen) or masked (hard screen). If the strategy chosen is the hard screen, these regions of the genome will not be assembled since overlaps cannot be computed. The second step is called overlapper. To find overlaps, Celera uses a method similar to BLAST. Each fragment is compared with all fragments previously examined. Overlaps are accepted if they have fewer than a certain percentage of differences and a minimum number of base pairs of unmasked sequences. Celera uses parallel processing in order to compare so many bases in a not too timely fashion. The fragments with a large number of overlaps are probably part of repetitive regions. The third step is called unitiger. Collections of fragments whose arrangement is uncontested by overlaps from other fragments are assembled into unitigs. If the unitig represents a unique sequence (as opposed to a repeat), it is called a U-unitig. Potential boundaries of repeat sequences are looked for at the ends of U-unitigs. When found, U-unitigs are extended as far as possible into a repeat. By detecting repeat boundaries, some overlaps between unitigs might be

resolved. The fourth step is called scaffolder. As its name indicates, all possible U-unitigs are linked into scaffolds which are sets of ordered and oriented contigs for which the size of the intervening gap is roughly known. When the two reads of a mate-pair are in different unitigs, their distance relation orients the two unitigs and allows to estimate the distance between them. Finally, the last step is the creation of a consensus sequence based on the different scaffolds.

4.2.5 Arachne

Arachne is used to assemble a whole genome [**?**]. It, too, is an overlap based algorithm. The first step is to detect overlaps and align them. The program identifies all k-mers ($k = 24$) and merges overlapping shared k-mers, then extends these shared k-mers to alignments and finally refines the alignment by means of dynamic programming. Arachne tries to achieve high-quality overlaps by correcting them before starting to assemble them. Once the overlaps have been identified, sequencing errors are detected and corrected by generating multiple alignments among overlapping reads using a majority rule based on the quality based score given by Phred. The alignments are then given a penalty score which combines individual differences among base calls. If the penalty score is too high, the alignment is discarded. At this level, repeats and chimeric reads are detected. The last step before contig assembly starts is identification of mate pairs. During the contig assembly, potential repeat regions are identified by aligning fragments that extend the same fragment. All fragments are merged and extended until a repeat region is found. When the contigs are assembled, Arachne goes back and

detects contigs that are potentially wrong due to repeats by looking at the depth of coverage and the consistency of linking with other contigs. Those contigs are marked. Once this step is completed, the software builds supercontig by incrementally using unmarked contigs. Finally, when all unmarked contigs have been assembled, Arachne tries to fill the gaps by using the repeat contigs.

4.2.6 EULER

EULER is an assembler based on a graph approach as opposed to the overlap layout consensus. This technique was developed to assemble reads obtained by sequencing-by-hybridization [**?**, **?**]. Let's say we want to reconstruct a sequence ATAGCATGCTT and the SBH gives us reads of length three. Those reads would be ATA, TAG, AGC, GCA, CAT, ATG, TGC, GCT, CTT. The reads are represented by nodes augmented with a directed edge between a node that has a suffix which is also the prefix of another node (for example, between ATA and TAG). In such a graph, assembling the reads would be equivalent to finding a Hamiltonian path. Since this problem is NP-complete, this construction has been discarded. Instead, a de Bruijn graph is built. With a de Bruijn graph, each $k-1$-mer is a node and there is a directed edge between two nodes N_1 to N_2 when there is an instance of a probe whose prefix is of a size $k-1$ is N_1 and whose suffix is of a size $k-1$ is N_2. This time, assembling the sequence is equivalent to finding an Eulerian tour in this graph.

This approach [**?**] is very close to that of EULER but EULER has additional modifications to it. First, before computing the eulerian path, EULER tries to

correct as many errors in the reads as possible. Indeed, each erroneous fragment will add wrong edges in the graph making it harder to compute the eulerian path. Also, EULER doesn't solve the Eulerian path problem but the Eulerian superpath problem. This problem is as follows; given an Eulerian graph and a collection of paths in this graph, find an Eulerian path that contains all paths as subpaths. To solve this problem, the graph created in the first step needs to be slightly transformed. Some improvements of EULER also use mate-pair, trying to solve repeats by treating each clone-mate pair as artificial paths in the graph with their expected lengths.

4.2.7 SOAPdenovo

The assemblers we reviewed previously were mostly based on long reads. In those cases, the overlap layout consensus approach makes sense but when the size of the reads is small, this approach starts to be less useful by itself and mixing it or using it with a graph approach (as with EULER) is probably a better choice. We start our discussion of assemblers for short reads with SOAPdenovo [**?**]. Before the program starts to assemble anything, there is a first step of preprocessing for error correction. For a small data set, this step is not necessary since the erroneous connections can be easily removed in the graph during the assembly. However, with large data sets (such as a human genome), this step might be crucial in terms of memory usage. Without it, the list of all reads (not cleaned of its errors) might be far too big to store in a machine's memory making the building of the de Bruijn graph impossible. Once this error correction step is done,

SOAPdenovo starts to assemble contigs. The initial graph is usually composed of 25-mers as nodes and the edge connection is made up of read paths. The tips that have a length smaller than a certain threshold are eroded in the graph. The assembler removes bubbles with an algorithm like Velvet's tour bus, with higher read coverage determining the surviving path. After the contigs are sequenced, SOAPdenovo realigns the reads onto the contigs. Each short read is mapped to one and only one contig without uncertainty since the repeat copies have been merged into consensus sequences in the graph and in the output contigs. The relationship between the contigs is then displayed as a graph. When repeat contigs have a conflict with the unique ones, they are masked. The remaining contigs with compatible connections are made into a scaffold. To join contigs into the scaffold, the information of mated-pairs is used.The last step is gap closure. Most of the gaps are due to the repeat contigs that were masked in the previous phase. To fill in the gaps, the paired-end information is used to get the read pairs where one of the reads is well aligned on the contigs and the other one located in the gap region.

4.2.8 AllPaths

Allpaths [**?**] is an algorithm that assembles microreads and paired reads. It starts by computing an approximation of the unipaths. A unipath is a sequence of nodes $x_1, \ldots, x_n$ in a de Bruijn graph for which $x_1, \ldots, x_{n-1}$ has outdegree one and $x_2, \ldots, x_n$ has indegree one and cannot be lengthened without violating one of those conditions. When the unipaths are computed, the first step is to chose

seeds. A seed is a unipath around which the sequence will be assembled. To pick those seeds, Allpaths looks for ideal unipaths which are relatively long with as low a copy number as possible (ideally one). Allpaths also looks at the pair reads information in order to spread those seeds as evenly as possible along the genome. After the seeds are picked, the assembler starts to build neighborhoods around them. A neighborhood of a seed is a region that extends the seed by 10 kb on each side of the seed. To construct this neighbor, the algorithm first finds a set of unipaths that partially cover the neighborhood. Then, two sets of reads are constructed, one composed of reads whose true genomic locations are near the seed, the other one made of all the short fragment read pairs near the seed. With the help of those two sets, the gaps between the unipaths of the neighborhood region are filled. The next step is to calculate the closures of all the merged short fragment pairs. The resulting set of closure sequences should cover the entire neighborhood region. Now, the only remaining local step is to glue together the closures of the mid-length read pairs. This gluing induces the assembly graph for the neighborhood. The local gluing runs in parallel and when this step is finished, Allpaths will build the global assembly. Basically, all the local neighbors are glued together, inducing a single sequence graph. This graph may have more than one component, depending on the number of chromosomes in the genome and also on the quality of the assembly. There is one last post-processing step in order to improve the quality of this graph.

4.2.9 Abyss

Abyss [?] is another assembler that works with short read sequences. The main structure in this algorithm is a de Bruijn graph, the originality here being the way the graph is implemented. Adjacent sequences do not need to be located in the same computer, allowing the program to distribute the sequences over a cluster of computers. The location of a given k-mer must be deterministically computable from its sequence. Also, the adjacency information between k-mers have to be stored independently of the location of the k-mer. The algorithm works in two steps. The first step is to build this specific de Bruijn graph, first spreading the sequences over the cluster then storing their adjacency information. Once this is done, vertices are not merged into contigs yet, but there is a run of read correction errors. When this cleaning is complete, the algorithm merges the vertices linked by unambiguous edges. Ambiguous edges are simply removed from the graph and the vertices are then merged creating the initial contig. This step closes the first phase of the algorithm. The second phase is to use the paired-end information in order to resolve ambiguities between contigs. This information is used to determine contigs that can be linked together.

4.2.10 SUTTA

In contrast to traditional graph based assemblers, a new sequence assembly method has been more recently developed. It employs combinatorial optimization techniques typically used for other well-known hard problems (satisfiability problem, traveling salesman problem, etc.). At a high level, SUTTA's framework

views the assembly problem simply as that of constrained optimization: it relies on a rather simple and easily verifiable definition of feasible solutions as "consistent layouts". It generates potentially all possible consistent layouts, organizing them as paths in a "double-tree" structure, rooted at a randomly selected "seed" read. A path is progressively evaluated in terms of an optimality criteria, encoded by a set of score functions based on the set of overlaps along the lay-out. This strategy enables the algorithm to concurrently assemble and check the validity of the lay-outs (with respect to various long-range information) through well-chosen constraint-related penalty functions. Complexity and scalability problems are addressed by pruning most of the implausible lay-outs, using a branch-and-bound scheme. Ambiguities, resulting from repeats or haplotypic dissimilarities, may occasionally delay immediate pruning, forcing the algorithm to lookahead, but in practice, do not exact a high price in computational complexity of the algorithm.

Chapter 5

SMASH

As we have seen in the previous chapter, sequencing whole genomes has been around for three decades and has gone through multiple innovations. Since Sanger, a number of new approaches have been created to form the so-called "Next Generation Sequencing". The goal of those new methods was to reduce the cost (in time and money) of the sequencing process compared to the Sanger method. Unfortunately, the current technologies and algorithms are not good enough to find rare SNPs or copy number polymorphisms. They simply ignore this problem. Those methods rely on aligners and assemblers that use shotgun assembly. It gives us a genotype consensus sequence but contain many gaps in the sequence which correspond to the repeats that we can find in a chromosome. The SNPs that we find using those technologies come only from non repetitive regions and they are haplotypically phased by using population data. Also, rare SNPs are rarely found. These technologies also force us to treat the Y chromosome separately and it is rather expensive. Finally, these technologies need bulk mate-

rials (a lot of cells) or amplifications which make them less useful for aneuploid cancer cells for example. Even when it has produced some form of a haplotype sequence (like Venter's), the sequencing requires a lot of post-processing operations, making the cost explode and the sequence still contains a lot of errors.

As we have discussed in the population genetics section of this document, we know that there is a need for a new sequencing technology and the priorities lie with an assembly algorithm that is cheaper and yet more accurate in producing haplotype sequence. The quality of a sequencing technology should not ultimately be assessed only on a base-by-base basis but also by the amount of information on genome structural information. It should be judged not only on an individual basis but on a haplotype basis.

How can one solve the problems we have just discussed? We can think of using a single molecule and a single cell. We will also need to have a long range sequencing technology in order to keep the context and be able to reconstruct a haplotype sequence. The major argument against this kind of approach is its high cost. One solution to this problem would be to use a hybrid technology. We could combine optical maps, Sanger sequencing and mate pairs in order to resolve our problems. This has been achieved by SUTTA [**?**]. Another approach and the one developed in this chapter is to integrate everything in one technology: SMASH (Single Molecule Approach to Sequencing by Hybridization). This method will reduce the errors and ambiguities of the resulting sequence while cutting down the cost. This technology combines other well-known technologies like optical maps and probe hybridization and ideas of SBH (Sequencing By Hybridization)

algorithms. The probes will give us short sequences and the optical maps will give us the context information necessary to obtain haplotype sequences. The caveat with SBH is its complexity but by combining those two technologies, we can tame this complexity.

We call SMASH-P the problem we are trying to solve and it can be formulated as follows. We are given a fragment (typically of length 4 kb) and a spectrum of this fragment. A spectrum is a map of all probes that are present within this fragment with their location information. With this information, we wish to determine the original sequence. Note that if one assumes that the single molecular data can be assembled into haplotypic maps, then at the end of our experiment we will have individual haplotype sequences. At a population level, that means we can have polymorphisms with exact phasing.

5.1 Sequencing Technology

We can separate the different sequencing technologies into two groups; those that focus on single base with an exact location of this base and another group based on long sentences without any location information. SMASH strikes a balance between those approaches. It is based on short words (k-mers or probes) with inexact location. This inexactitude gives us a window of a certain size where we can find our probe. The set of all probes with their associated locations is called a spectrum and with this spectrum, we are in a situation where we have to solve the positional SBH problem described in [**?**].

In practice, those windows allow us to treat our problem in a divide and con-

quer fashion. Each one of these windows is independent from the others and can therefore be treated separately. This approach makes our technology highly parallelizable. When we are dealing with haplotypic optical maps, these windows are nothing but the different restriction fragments given by the optical mapping technology (explained in more detail in section 5.1.1).

As we have seen, we will have to assemble our sequence for each of the restriction fragments. This assembly can be carried out independently so we can just focus on what happens for one of those fragments, the same reasoning being applicable to all the fragments. For each fragment, we get a spectrum (explained in more details in section 5.1.2)which is the set of all the probes present within this fragment with their location. Such a spectrum is corrupted with some noise which can be typically put into three groups: false positives, false negatives and location error. The simplest scheme is non robust because of the non random nature of a human genome. Places where we find repeats or certain type of patterns may pose difficulties for the algorithm. By introducing the use of universal bases, this limit can be ameliorated as show in [**?**], [**?**] and [**?**].

5.1.1 Optical Restriction Fragments Mapping

We want to create technologies that are accurate, inexpensive, flexible and produce whole genome haplotype sequences. Having the haplotype will permit later study on genomic variations at multiple scales and across multiple species. To develop such technologies, we can integrate components of technologies that are

used for various mapping approaches like optical mapping or array-mapping techniques. We can find a description of those in [?], [?], [?], [?], or [?]. The advantage of these techniques is that they can provide us powerful algorithmic strategies that may be capable of statistically combining disparate genomic information and novel chemical protocols that can, in parallel, manipulate and interrogate a large number of single DNA molecules in various environments.

Our sequencer can incorporate several of those technologies. One of these is a single molecule technology, often called Optical Mapping and described in [?] and [?]. Another optical mapping approach is based on an LNA/PNA probe technology that hybridizes to double-stranded DNA. Optical Mapping is a single molecule approach allowing us to detect genetic markers. Raw optical mapping can be assembled on computers in order to get whole genome haplotype restriction maps.

We can use Optical Mapping to build up single molecule DNA ordered restriction maps (also called physical maps) using fluorescent microscopy. We can find a description of this in [?] and [?]. After several years of work and effort spent on Optical Mapping, the first single molecule mapping technologies for BAC clones was released in 1998 in [?]. A year later, a technology based on the Gentig algorithm for whole microbial genomes was published in [?]. DNA is extracted directly from cells by lysing (without the use of clones). It can be sheared into 0.1-2Mb pieces and attached to a charged glass substrate. Then, a reaction occurs with the restriction enzyme and finally, DNA is stained with a fluorescent dye as described in [?]. The gaps created by the restriction enzyme can be spotted

with a fluorescent microscope and appear as breakages in the DNA.

The images collected by the microscope can be processed by imaging algorithms to detect the brightness of the molecule. It will also detect cleavages within the molecule, therefore detecting the restriction enzyme sites. The distance between such sites can be approximately estimated by comparing the integrated fluorescent intensity relative to that of a standard DNA fragment that has been added to the sample. Using the length and the restriction map of the standard, we can deduce the distance between sites in the studied molecule. Using a fluorescent probe that hybridizes at the end of the standard DNA makes it even more readable and recognizable in the image, improving the overall technology.

Obviously, errors can be introduced during the experiment and the analysis. The restriction enzyme may not cut the DNA at some sites. The DNA could randomly break, creating a gap that cannot be distinguished from a cleavage site. The dyeing process may not be homogenous. The image processing might make mistakes in detecting gaps (missing some real ones or creating new false ones). Those kinds of errors can be categorized in a raw map. We can face sizing errors in the fragment or the distance between two sites (of the order of 10% for a 30Kb fragment). Also, missing restriction sites can occur (10 to 20% of the restriction can be false negatives) or false restriction sites (2 to 10% of restriction sites can be false positives). Finally, we can have missing fragments (half of all fragments under 1Kb and most fragments under 0.4Kb). To recover from those errors, we can use redundant data. A minimum redundancy of 50x can be used to assemble genome wide maps and recover from most errors with high confidence, as de-

scribed in [?] and [?].

Even though optical mapping of whole organism genomes may be produced using conventional techniques as described in in [?], [?], [?] and [?], we want to employ those techniques in a different fashion. We utilize a restriction enzyme that will give us restriction fragments on an average size of 2-16kb and at least 50X coverage (50x for each haplotype) and will enable us to assemble a genome wide haplotype. This restriction fragment map will provide a scaffold for sequencing the genome.

5.1.2 Optical Probes Mapping

We hybridize fluorescent oligonucleotide probes to DNA. Various types of probes can be used as we will see. Fluorescent microscopy images of the hybridized DNA can be assembled by computers into genome wide haplotype maps of location of the probe sequences. The sizing information of that map will not be as accurate as a restriction map but by tallying up the same restriction sites to the molecules with the probe sites, the sizing can be normalized every 2-16Kb. This process can generate a map for any probe sequence using standard coverslips covered with genomic DNA using a molecular-combing-like technique for flow deposition of the DNA.

The cost for sequencing human whole haplotypic genome can be dominated by the cost to image standard 20x20mm regions on a fluorescent microscope at a resolution of 1 pixel every 75nm. A design for such a microscope system,

designed to minimize cost and maximizing throughput, is described in a proposal to NIH for a Novel Whole Genome Sequencing Technology by Anantharaman et al. in 2005 (never published) and may be based on conventional components that can image a large number of coverslips per day. There is also room to design customized fluorescent microscopes and VLSI chips for high throughput CD imaging to improve this technology in order to reduce the costs.

We wish to hybridize those probes with genomic DNA without breaking the DNA. We can deposit DNA intact on a surface, as for the restriction enzyme mapping technology. Regular oligonucleotide probes (as used in FISH, for example) will typically hybridize at 75°C. This temperature is above the melting point of dsDNA (double stranded DNA, which is typically 65°C). Hence, this treatment can result in breaking both strands of dsDNA and produce random "necklaces" of DNA balls (often seen in Fibre-FISH) instead of one continuous segment of DNA. Such a behavior can be seen in [**?**], [**?**] and [**?**]. Another problem with regular oligonucleotide probes is that the length of such a probe for a reliable hybridization should be of 15bp or longer. Fortunately, there are other types of probes that do not break dsDNA and that can hybridize reliably with only 6bp. Here is an overview of such probes.

LNA (locked Nucleic Acid) probes are single stranded, like PNA (Peptide Nucleic Acid) probe. The difference with PNAs is that they rely on a greater specificity to ssDNA (single stranded DNA). We can find a description of LNAs in [**?**] and [**?**]. The advantage of both LNAs and PNAs is that they can hybridize with dsDNA at 55°C and therefore will not break our molecule of dsDNA. At this

temperature, dsDNA will frequently open their two complementary ssDNA at various locations, allowing our LNAs or PNAs to hybridize. When a LNA probe (or PNA) hybridizes to ssDNA, it remains bound since its binding constant is higher than that of dsDNA. As mentioned before, LNA has a stronger affinity with ssDNA than PNA and depending of the GC content of the sequence, the length of the LNA that reliably hybridizes with DNA may vary from 6 to 8 bp as described in [**?**].

In contrast with LNA and PNA, TFO (Triplex Forming Oligonucleotide) probes can hybridize directly to dsDNA without having to "open" the DNA into two ssDNA. When it hybridizes with dsDNA, it forms a triple stranded DNA. TFOs have originally been developed for suppressing gene expression *in vivo* in [**?**] but can also be utilized as fluorescent probes. A common TFO design can be an oligo formed by a 50% mix of LNA and normal DNA. It can be improved employing ENA (Ethylene Nucleic Acids). The melting temperature for TFOs varies from 28°C-41°C for regular ones and 42°C-57°C for ENA-DNA mixtures.

Double stranded probes can be designed using pcPNA (pseudo-complementary PNA) which is a modified form of ssPNA probes that may not hybridize with themselves as shown in [**?**] and in [**?**]. Complementary pairs of such probes may be used to hybridize with both strands of the dsDNA, which can be stable because the two pcPNA-DNA hybrids formed may be more stable than dsDNA.

For this technology, after preliminary experiments with LNA probes, it was decided to keep pursuing the use of PNA probes and more precisely, bisPNA. To test the efficiency of hybridization of bisPNA, it was hybridized it to lambda

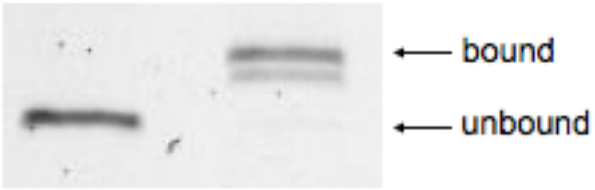

Figure 5.1: 880 bp fragment resolved using 4% PAGE gel. The first lane is the lambda DNA sample without bisPNA probe hybridization digested with PmlI restriction enzyme. The second lane is the lambda DNA sample that has been hybridized with bisPNA probe digested with PmlI restriction enzyme. There is a clear shift in mobility of the 880bp fragments, which has bound the bisPNA probe.

DNA molecules inside a test tube. The probe target was an 8-mer sequence (5-GAGAAGGA-3). To measure the quality of the hybridization of this probe, the lambda DNA was digested after the supposed hybridization with PmII restriction enzyme and run the sample on a 4% PAGE gel. It was found that the rate of hybridization was greater than 90%.

5.1.3 Results

The focus was on two kinds of tests. Mishra-lab started with small genomes like E. Coli to keep the experimental cost low. The goal of this experiment was to validate the scheme of using a combination of restriction and probe maps and also to estimate various parameters. The goal was to achieve restriction enzyme mapping and probes hybridization mapping simultaneously. The digestion of a molecule by a restriction enzyme had an efficiency of the order of 90%. At the same time, hybridization had an efficiency of only about 30%.

When one examines the image, only 30% of the matching probe sites are

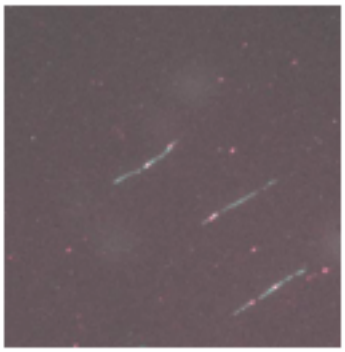

Figure 5.2: Overlayed fluorescent images of labmda DNA molecules using a FITC filter (white) and CY5 filter (red), showing the position of the probes on the lambda DNA molecules.

visible. One must ensure that, to assemble genome wide maps from restriction fragments, the false negative rate should not exceed around 30% per marker site as shown in [?]. It follows a 0-1 law. If experiments operate above those parameters, it can produce reliable maps. One can get a likely false negative rate of 70% for probe maps by carefully setting up the experiment in this way.

The scientists in Mishra-lab used a suitable threshold to minimize false positives. They then estimated the distance between probe locations (or the DNA ends) by comparing the intensities of the two images. The resulting probe map from each DNA molecule is normalized to the same length of 100%. The most likely consensus map was computed by combining probe maps from around 20 image pairs using a Bayesian algorithm. For one set of 20 image pairs, a total of 512 DNA molecules with a total of 678 probes were identified and combined into a consensus map with 2 probe locations at 14.8% and 52.4% of the DNA length. The 3' to 5' orientation of the DNA molecule cannot be determined from optical

maps. Thus this result is in close agreement with the correct map with probes at 50.2% and 85.7% (14.8% $\approx$ 100% $-$ 85.7%). The probe hybridization rate of 42% is also quite good.

They next generated a high resolution ordered E.coli K-12 genome map using both hybridizing probes and an XhoI restriction digest of single DNA molecules. The K-12 bisPNA probe was designed to target a specific 8-mer sequence (GAAGA-GAA), which appear 313 times along E.coli K-12. They used the same fluorescent hybridization technique that was used in the creation of the lambda DNA map. Separately, they digested the labeled single DNA molecules with XhoI restriction enzyme and combined the mapping information from both approaches.

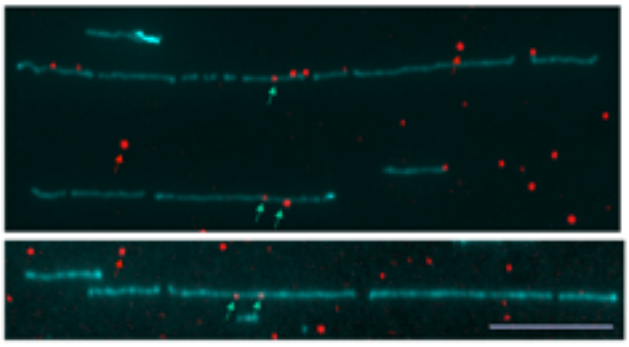

Figure 5.3: Experiments with E. coli K-12 genome.

The initial results showed successful generalization of this technique, initially developed to map lambda DNA. Thus it was seen that it is possible to combine optical mapping and hybridization.

5.2 Assembler Algorithm

We will now introduce the algorithm by Anantharaman, Lim, Mishra (unpublished results). For now, we will only focus on a restriction fragment of the sequence since we have seen that we need to solve the same problem for every fragment. At the end of the experiment described in the first part, we end up with a probe map or positional spectrum which is the set of all possible L-mers with their locations. Ideally, the information generated by restriction digestion and sequencing of probes would consist of a triplet of locating data for every possible probe generated by the restriction enzyme digestion:

- sequence (5' to 3') of the template (or expressed) strand,
- sequence (5' to 3') of the complementary strand, and
- position (or positions, if a sequence appears more than once) (number of base pairs from 5' end) of the 5' end of each sequence; template and complementary.

In short, a triple of the map is of the form (x, ω^W, ω^C) where x is the position of the probe, ω^W the sequence of the probe in the template strand and ω^C the sequence in the complementary strand. The goal of the assembly algorithm is, from this positional spectrum, to construct a sequence τ that is coherent with the given map. We can make an analogy with trying to read a book from an index. In the index of the book, all the words are referenced with their page, line and position in the line numbers.

If all three of these factors could be entered with high accuracy, generating a sequence would be a straightforward matter. Such a world does not exist and so we have to face data with errors of different kinds. We need to take this noise

into account if we want our sequence τ to be the same as our sequence σ.

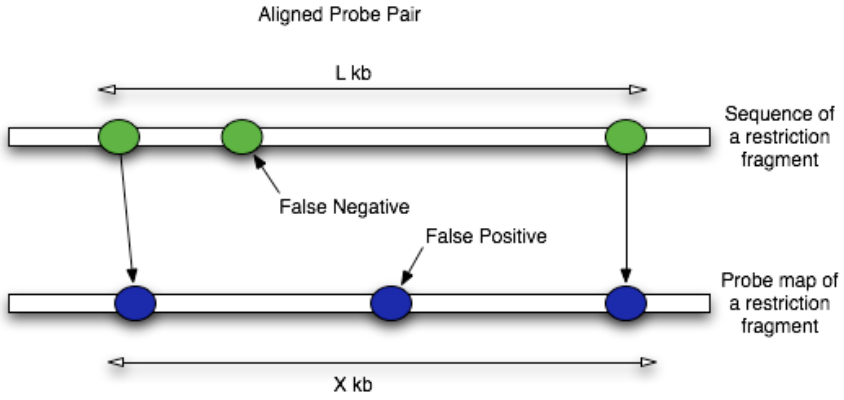

Figure 5.4: For the restriction fragment of the DNA we are currently treating (usually of length 1kb), we can see the different types of noise. In green are some probes along the sequence. We can see that the second green probe does not appear in the positional spectrum (here, the spectrum is represented as if it were already reconstructed in a sequence) and so is a false negative. We also see that the first green probe is a match with the first blue probe with a small shift (location error). Finally, the second blue probe, used to reconstruct the sequence, does not appear in the original sequence and so is a false positive.

We can divide the noise into three different components. The first is the location error. A probe that has a location error is a probe that represents the reality of the sequence we are sequencing but that is slightly shifted from its real location by a window of a certain size in bases. Another type of noise is false positives. A false positive is an L-mer that is present in the map but absent in the original DNA sequence. Typically, a false positive probe is a probe that is shifted by more than the accepted window size. Finally, we also have to deal with false negatives, the opposite of false positives. A false negative is an L-mer that is present in the original DNA sequence but not in the map. If we come back to our book analogy, we now have to read a book from an index that contains

words that are not present in the book (false positives), that misses some of the word that are in the book (false negatives) and some words are referenced with a wrong number of page for example (the location error).

Now that we have a model for our noise, we can assemble the map into a sequence. There are 5 basic steps, each of which will be described below:

- Start with a sequence of $k-1$ bases (this sequence can be derived in various ways).
- At the kth position, add each of the 4 possible bases, and score the probability of the k, using the map as a guide.
- At the $k+1$ position, repeat step 2, then, add the scores of k and $k+1$ for each possible sequence. Repeat for each subsequent base. A tetranary (base 4, as there are four possible bases at each position) tree is formed.
- Prune the tree occasionally, removing the sequences with the lowest scores
- Repeat until the false negative rate jumps from 2% to 55%.
- Choose the sequence with the best score.

Initial $k-1$ sequence: For software testing purposes, the initial $k-1$ sequence of base pairs can be determined from the reference sequence (which has been artificially digested to create a map), though in an actual sequencing situation, all possible $k-1$ probes must be created. The incorrect probes will quickly get pruned as the sequence grows past the first few bases. Because all the probes on the actual positional spectrum are k bases long, it is impossible to score a probe of the first $k-1$ bases alone, since all scores must be based on the probability of a probe of k bases.

Adding the kth base: At the kth position, all 4 bases are added to each of the constructed initial probes. Because each probe is now k bases long, they can be compared to the map. Based on map-reported probe sequences for the first k positions, a score is assigned to each of the computer-generated probes

Adding Subsequent Bases: At the $(k+1)$th position, all four bases are added to each leaves of the previous tree (of depth k). Again, a score is generated for each of these new probes based on map-reported probe sequences for positions 2 to $k+1$. This score is added to the score generated for that same sequence score for position k. We then iterate this operation as many times as necessary to finally reconstruct the entire sequence. Obviously, the tree can grow exponentially and must be pruned regularly.

Pruning the tree: The sequence assembly heuristic described above can be achieved in linear time because it is possible to limit the number of paths at any depth of the tree to some maximum number (which can be referred to as the beam width). Whenever the number of paths exceeds this maximum number, a sufficient number of worst scoring paths can be discarded such that the remaining number of paths drops below the beam width. There can be a small risk that the correct path (which may not be a best scoring path) may be discarded too hastily. Simulations indicate that for random sequences, such an early discarding of the correct path may not occur if the beam width is set to the equivalent of 2 Gigabytes of memory. For a human genome sequence, the correct sequence may be discarded about once every 50kb. Even in such cases, the incorrect sequence assembled may be usually incorrect only in a few bases

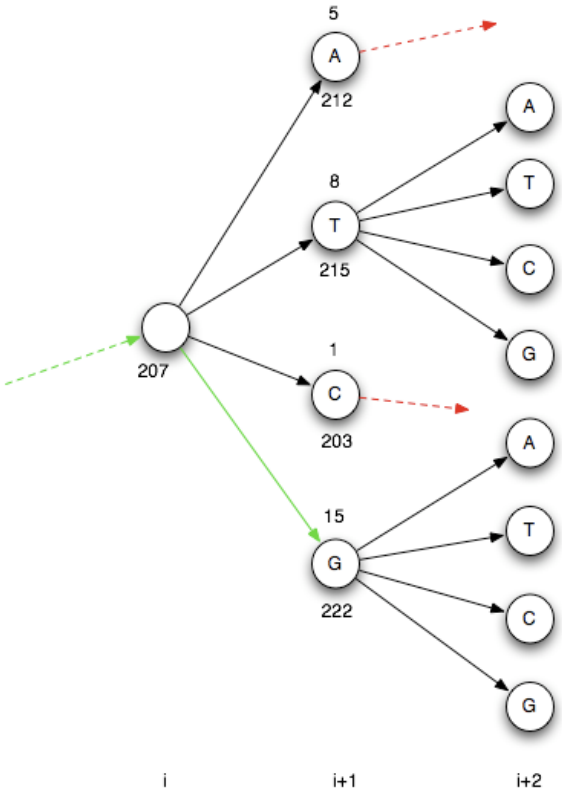

Figure 5.5: The first i positions of the sequence have already been computed. At position i, we add the 4 possible bases. We compute a score for each of the bases (upper number). The score for the sequence of length i+1 is the score of the sequence of length i + the score to add one of the bases to this sequence (lower number). If the number of paths exceeds the beam width, the worst paths (in term of score) are pruned (the red dashed arrows) until we have reached a number of paths below our beam width. The green arrow represents the best scoring path.

(typically 10-30bp) around the region where the beam width was exceeded. Such errors can be reduced further, by adding an annealing step in which regions of the assembled sequence that are likely to contain errors (e.g., regions where the beam width was exceeded) may be subsequently reassembled locally while relying on the higher level of correctness of the sequence on either side of the problem region.

5.2.1 Results

Gapped Versus Ungapped Probes

We wanted to create simulated data from real human genome and check the algorithm for two different approaches, one with ungapped probes and one with gapped probes (use of universal bases). To generate the simulated data we used both random DNA sequences as well as sequences from *H. sapiens* chromosome 1 and computed the probe map of a single restriction fragment of size 1kb, for all possible probes for the probe type chosen. For example, for a probe with 6 specific bases and 4 universal bases and the pattern xx-x–x-xx (x being a solid base and a dash a universal one), there are a total of 2080 distinct possible probes, excluding reverse complements. For each probe map, we simulated data error under the following assumptions for single DNA molecules: Probe location Standard Deviation = 240 bases; Data coverage per probe map = 50x; Probe hybridization rate = 30%, and false positive rate of 10 probes per megabase, uniformly distributed. Instead of simulating each single DNA molecule, we analytically estimated the average error rate in the probe consensus map based on the above assumptions: Probe location SD = 60 bases; False Positive rate $< 2.4\%$; False Negative rate $< 2.0\%$. Using these estimated error rates for probe consensus maps we randomly introduced errors at the above rates into each of the 2080 simulated probe consensus maps (for the above example). We then ran our sequence assembly algorithm, and then aligned the sequence produced with the originally assumed

correct sequence using Smith-Waterman alignment. We counted the total number of single base errors (mismatches + deletions + insertions). We then repeated this experiment until a total of 200,000 bases of sequence had been simulated and computed the average error rate per 10,000 bases. We first tried probes without universal bases with 5,6,7 and 8 bases respectively and got error rates per 10,000 bases of 1674, 255, 39.6 and 3.7 bases respectively.

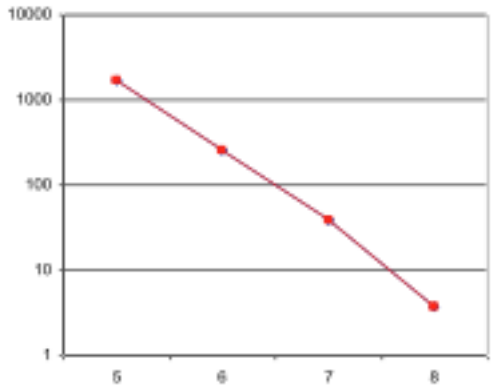

Figure 5.6: Sequencing errors per 10kb sequence for solid (no universal bases) probes

Next we tried various gapped probes (with universal bases) each with 6 specific (solid) bases and varying the numbers of gapped (universal) bases, ranging from 1 to 5. We always put 2 solid bases at each end and placed the remaining two solid bases so that the resulting pattern was symmetric, since that ensures that there will only be 2080 distinct possible probes (rather than 4096 possible probes for non-symmetric patterns of solid bases). The exact patterns used were xxx-xxx, xx-xx-xx, xx-x-x-xx, xx-x–x-xx, and xx–x-x–xx respectively. The resulting errors rates per 10,000 bases with 1,2,3,4 and 5 gapped probes were 35.9, 4.35, 2.65,

0.05 and 0.30 respectively. We excluded regions within 5 bases of a simulated restriction site, since error rates are higher at those locations.

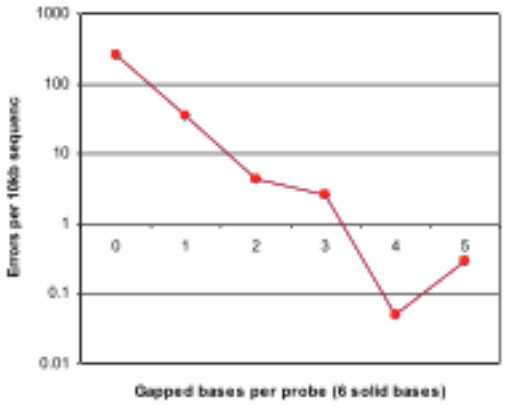

Figure 5.7: Sequencing errors per 10kb sequence for gapped probes

Note that while the error rates mostly decreased monotonically as the total probe size increased, the probe with 5 gapped bases had a higher error rate than the one with 4 gapped bases. One possible explanation is that the patterns chosen are not optimal, and in particular the 5 gap pattern is less optimal than the 4 gap pattern. We have subsequently explored additional patterns to determine the optimal gap pattern, which has made it clear that the probes with 4 and 5 gap bases far exceed the goal of 1 base error per 10,000 bases as desired in applications involving rare and de novo mutations. Note also that the error rates of gapped and ungapped probes of the same length roughly match for lengths of 7 and 8 bases, in accordance with the theory for optimal probe patterns, suggesting that the patterns we picked for 1 and 2 gapped bases are already close to optimal.

FN (%)	% of correct assembly
0	97.48
0.5	97.79
1	97.70
1.5	97.59
2	97.87
2.5	97.60
3	97.43

Table 5.1: Percentage of sequence correctly assembled for different values of false negatives while other parameters (false positives, window error size, probe pattern) vary

Robustness To Parameters

Considering that gapped probes produced better results, we then changed the parameters of our simulations to see how robust the algorithm was. We made the probe location window vary from 0 to 105 bp (0% to 10.5% of our fragment size) by increments of 15. We also tweaked the false positive and false negative rates from 0 to 3% by increments of 0.5%. We focused on 15-mers with 6 solid base pairs and therefore 9 universal bases. We reconstructed 20 kbp of the chromosome 1 sequence. So for example, for 1.5% of false positives, we will get the result of the experiments with 1.5% false negatives and all the values of the other parameters (6 for the false negatives, 7 for the sizing error, 15 for the pattern and 20 for the size of the sequence). That gives us 12 600 experiments on which we compute the average score of the alignment between our assembled sequence and the reference sequence. The final percentage we get is the percentage of the sequence that is correctly assembled. For a 97% result, that means on a sequence of 100 bp, we have made 3 mistakes.

FP (%)	% of correct assembly
0	97.93
0.5	97.52
1	97.59
1.5	97.73
2	97.60
2.5	97.60
3	97.49

Table 5.2: Percentage of sequence correctly assembled for different values of false positives while other parameters (false negatives, window error size, probe pattern) vary

We notice in Table 1 and 2 that the percentage of false negatives or positives does not have any effect on the result, which means that our algorithm can handle a reasonable amount of these kinds of noise without a problem. On the other hand, we see in Table 3 that as the sizing error grows, the quality of the assembler diminishes and the closer we get to 10% of the length of the fragment (we reconstruct 1 kbp fragments so 10% is 100 bp), the more inaccurate our algorithm is, as we have discussed earlier. Finally, we see that the choice of a pattern is fairly robust since only few of them (3 over 15) are significantly worse than the others. We can also see that the values of percentage of sequence correctly assembled are sometimes low (around 3% of mistakes for the false positives or false negatives rates). Our goal here was to get an idea of what pattern is good or to know if the value of a parameter has any effect on the execution of the algorithm. This requires a lot of simulation so we decreased the number of branches saved in our tree to execute the simulations faster. As we prune more branches, the risk of mistakes becomes higher and therefore, we have more mistakes than if we were to run the algorithm normally.

Location error (bp)	% of correct assembly
0	99.22
15	98.60
30	98.40
45	98.43
60	98.29
75	97.43
90	96.40
105	94.32

Table 5.3: Percentage of sequence correctly assembled for different values of sizing errors while other parameters (false negatives, false positives, probe pattern) vary

It is clear that the sizing error should be controlled since we have a large decrease in accuracy as the value of this parameter gets closer to 10% of the fragment length. However, the rate of false positives or negatives does not significantly impact the execution of our algorithm (except for the time of execution) for at least 3 % which is a reasonable value for a real life experiment. Finally, choosing the right probe design may be important in order to have the best assembly possible. It will be interesting to see if there is a combinatorial structure behind the "good" patterns and the "bad" ones so we could predict in advance what pattern we should design before starting the experiment.

5.2.2 Complications

There may be repeated regions in a sequence leading to wrong paths that look correct. Every time we hit one of those regions the number of such paths will keep multiplying and might make our tree grow exponentially. Fortunately, this situation can be avoided. We can label each probe in the map with its multiplic-

Probe pattern	% of correct % assembly
$x-x-x-----x-x-x$	91.69
$x-x---x-x---x-x$	91.92
$x---x-x-x-x---x$	92.24
$x---xx---xx---x$	97.88
$x--x--x-x--x--x$	98.47
$x--x-x---x-x--x$	98.75
$x--xx-----xx--x$	98.77
$x----xx-xx----x$	98.88
$xx---x---x---xx$	98.99
$xxx----------xxx$	99.12
$xx-x---------x-xx$	99.13
$xx--x-----x--xx$	99.21
$x-xx---------xx-x$	99.23
$xx----x-x----xx$	99.29
$x-x--x---x--x-x$	99.58

Table 5.4: Percentage of sequence correctly assembled for different probe patterns while other parameters (false negatives, false positives, window error size) vary

ity depending on the intensity of the fluorescence we observe in the microscope. Then you can penalize a path in the graph that uses a probe that has already been used as many times as its multiplicity. That would avoid a case where we assemble too many repeats. On the other hand, any final sequences not containing enough repeats to explain the multiplicity of certain probes can be penalized. This penalization requires looking back to count how many times a probe has been used. This step can be very slow even if going back just to the previous occurrence of the probe is sufficient (this occurrence can be thousands of base pairs away), if it needs to be done every time the path is extended by one base pair. To prevent this issue, we use two types of data structures. One is a table containing the probe location at selected nodes in the tree. At those nodes, the table contains the previous location of each probe. We store this table every 64 nodes which limits the amount of memory per node (130 bytes per node for 6-mers probes, and this value can even be lowered). To find the first instance of the probe, we look back to one of those "special" nodes. Finally, in order to find the remaining locations of the probe in the path, we add a pointer that refers to the previous node that has the similar probe instance as the current node. Hence, we only look back at 64 plus the number of occurrences of a probe nodes instead of the thousands previously described.

There are other types of structures that we can find in the genome which lead to problems in reassembly. One of those is when we have a sequence following this form: $xW\bar{x}$ with $\bar{x}$ representing the reverse complement of x. During the

execution of our algorithm, there is a risk that we will reconstruct $x\bar{W}\bar{x}$ instead of $xW\bar{x}$. As an example, consider the following DNA sequence:

TATCACCGGATA (W)

ATAGTGGCCTAT (C)

We see that GATA is the reverse complement of TATC (here, W and C stand for the Watson and the Crick branches). Assuming we use 3-mers, the probe map that we would obtain for such a sequence would look like TAT, ATC, CCG, CGG, GAT, ATA, TCA, CAC, ACC, GGA, TCC, GGT, GTG, TGA. The underlined probes are those where our algorithm will not be able to determine which probe to use to continue assembling. If we try to reconstruct this sequence by hand, here is one possible result:

W	T	A	T									
W		A	T	C								
C			T	C	C							
C				C	C	G						
C					C	G	G					
C						G	G	T				
C							G	T	G			
C								T	G	A		
C									G	A	T	
C										A	T	A
	T	A	T	C	C	G	G	T	G	A	T	A

We started by assembling the Watson branch but when we hit the red probes, because they are both almost equally plausible, one had been chosen randomly, leading to the overall reconstruction of our Watson branch. If we go back to our graph representation, that kind of structure would be a cycle in the De Bruijn graph but we would not know what direction to enter the cycle. By using 6-mers, we fail to see far enough ahead in this loop to assess what is the correct direction to enter the cycle. Interestingly, such structures are few in the human genome (in the order of 50bp).

Suppose we are looking for a pattern $Pat = A\bar{(}A, \bar{B})^i B$. Let us define the probabilities $p(A) = p(B) = p$ and $p(\bar{A}, \bar{B}) = q = 1 - 2p$. The probability that such a pattern Pat will occur in a sequence of size L will be $p(Pat) = \sum_{i=0}^{L} p^2 q^i = p^2(1 - q^{L+1})/(1 - q) = (p/2)[1 - (1 - 2p)^{L+1}] \approx (p/2) * (2p(L + 1)) \approx p^2 L$. This

approximation will give us the expected number of pattern Pat occurring in a restriction fragment of size R. This number is $E = p^2 * LR$ with $p = 1/4^k$ for any k-mer.

Now if we look at the value of E depending on the size of our k-mer, we can see that for a 6-mer, and $L = 4000, R = 50, k = 5$, we have $E = 0.191$ but for an 8-mer, with $L = 4000, R = 50, k = 7$, we have $E = 7.4 * 10^{-4}$. Even better, for a 15-mer with $L = 4000, R = 50, k = 14$, we have $E = 2.78 * 10^{-12}$.

We can see here that with 6-mers, this kind of pattern can be expected relatively frequently but going to 8-mers (and even more with 15-mers) would actually solve the problem for the vast majority of the cases. On the other hand, with 6-mers probes, only 2080 experiments need to be performed while with 8-mers, 32896 experiments are necessary, significantly increasing the cost of the technology. We would like to keep the number of experiments as low as possible but we need to overcome the problems encountered with those small probes.

One efficient way of confronting this problem is to use so-called universal bases. Universal bases are bases that can bond to any of the A, T, C or G base. Their efficiency has been discussed in [**?**]. We can now use probes that are longer (say 15 base pairs) but that are still made of 6 solid base pairs and therefore 9 universal bases (e.g ATT—G—C—CCT would be such a probe, with the dash symbolizing a universal base). Obviously, by using those "wild cards", we lose a little bit in terms of accuracy but it does not create a real problem. On the other hand, we are now able to have 15-mers for the same cost as 6-mers.

5.3 Improvements

5.3.1 Design of gapped probes

We can see here that even though most of the gapped probes give us good results, some of those probes are not as good as the others. Even if the algorithm seems pretty robust to such a choice, choosing the optimal pattern for a probe might lead to easier reconstruction of our sequence and therefore faster and more accurate results. It would be interesting to see if we can predict a priori which patterns are better.

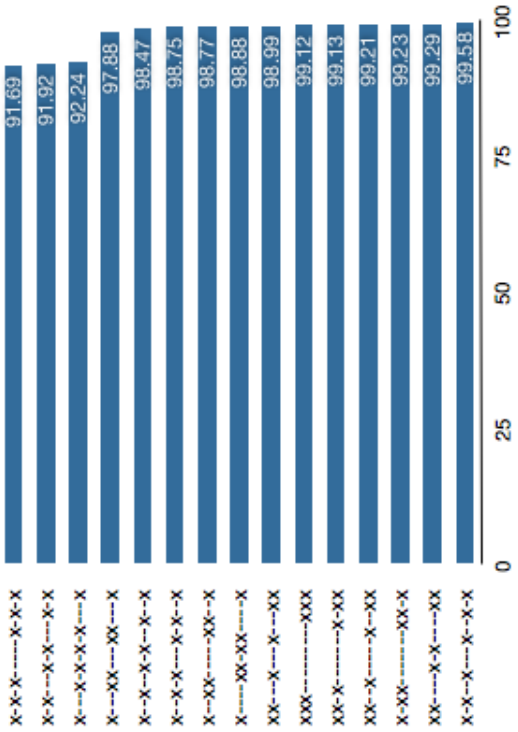

Figure 5.8: Percentage of correct assembly of our sequence for different probe patterns.

				i			
x	x	x	x				
	x	x	x	x			
		x	x	x	x		
			x	x	x	x	

Table 5.5: Coverage of position i by ungapped probes of size 4

					i																		
x	x	-	-	x	x							x	-	x	x	-	x						
	x	x	-	-	x	x							x	-	x	x	-	x					
		x	x	-	-	x	x							x	-	x	x	-	x				
			x	x	-	-	x	x							x	-	x	x	-	x			
				x	x	-	-	x	x							x	-	x	x	-	x		
					x	x	-	-	x	x							x	-	x	x	-	x	

Table 5.6: Coverage of position i by two different gapped probes of size 4 with two universal bases

When we use ungapped probes, a specific position i of the sequence is going to be covered by k different probes (if the size of the probe is k) and the coverage of the surrounding $2k-2$ is going to be maximal.

Now we have gapped probes that lose this maximum coverage. Each position is still covered by k probes but the size of the probe includes both solid and universal bases. Depending on the probe pattern, the coverage is different.

The reason we have good and bad patterns might lie in this difference of coverage. Some probe patterns might allow a better mixing of solid and universal bases. We can represent the different coverages of the probes by their graph and

adjacency matrices. For example, the matrix for the right pattern of Table VI is:

$$\begin{bmatrix} 6 & 1 & 2 & 2 & 0 & 1 \\ 1 & 6 & 1 & 2 & 2 & 0 \\ 2 & 1 & 6 & 1 & 2 & 2 \\ 2 & 2 & 1 & 6 & 1 & 2 \\ 0 & 2 & 2 & 1 & 6 & 1 \\ 1 & 0 & 2 & 2 & 1 & 6 \end{bmatrix}$$

A good indicator of how well our probes are mixing is the value of the spectral gap of those matrices. The spectral gap is the difference between the first and the second eigenvalue of a matrix. If the spectral gap is big, it is easier to go through our graph from any point while a small spectral gap means that it is hard to travel in the graph. What that means for our purpose is that if our probe pattern has a big spectral gap, even though the probe is gapped, the coverage of the few bases around the currently treated base is good enough to give us some checking information.

We normalize our matrices so they are stochastic and the first eigenvalue will be 1 for each of them. We now compute the second eigenvalue of the matrices. Here are the results:

We notice two effects here. First, the three probes that were not good to assemble the sequence get a second eigenvalue of 1, leading to a null spectral gap. By looking at the eigenvalues, it seems that we are able to predict the clusters of good and bad probes. Unfortunately, we are not guaranteed to get

Probe pattern	% of correct assembly	2nd eigenvalue	spectral gap
$x-x-x-------x-x-x$	91.69	1	0
$x-x---x-x---x-x$	91.92	1	0
$x---x-x-x-x---x$	92.24	1	0
$x---xx---xx---x$	97.88	0.7223709	0.2776291
$x--x--x-x--x--x$	98.47	0.7262696	0.2737304
$x--x-x---x-x--x$	98.75	0.5002489	0.4997511
$x--xx-------xx--x$	98.77	0.5691417	0.4308583
$x----xx-xx----x$	98.88	0.4412064	0.5587936
$xx---x---x---xx$	98.99	0.7096873	0.2903127
$xxx---------xxx$	99.12	0.8679781	0.1320219
$xx-x---------x-xx$	99.13	0.6529811	0.3470189
$xx--x-------x--xx$	99.21	0.5141738	0.4858262
$x-xx---------xx-x$	99.23	0.6116911	0.3883089
$xx----x-x----xx$	99.29	0.6261011	0.3738989
$x-x--x---x--x-x$	99.58	0.6202461	0.37975389

Table 5.7: Value of the spectral gap for the different 6-mers

the optimal probe. This does not pose too much of a problem since the precision in the assembly process for the good patterns are fairly close and would be even closer if we had simulated the assembly with a bigger memory.

Conclusion

Ten years ago, when the Human Genome Project started, the hopes were tremendous and the expectations were high. A decade later, we find ourselves in front of a door which beckons an ambiguous future. Will this door open to a new era in term of medicine and biology discovery or will it close and remain closed to hide a major failure? Newer and newer sequencing technologies are being developed and improved but many people still doubt if these technologies will yield any useful results. Genome-wide association studies have reached a dead end. While new technologies have been focusing on cutting costs and increasing throughput, they have lost accuracy, allowing for more single nucleotides and indel errors. Worse, they still cannot sequence haplotypes. Despite these issues, we feel that there is hope for the future or genome-wide association studies. Overcoming these difficulties requires the development and design of a highly performing technology that is able to sequence haplotypes with an acceptable rate of mistakes and still operate at a reasonable price. With this technology in hand, the study of populations may become more efficient and could lead to results that will live up to the expectations biologists and doctors once had.

This dissertation has presented solutions to those problems. A new sequencing technology called SMASH has been introduced. The combination of two technologies utilized by SMASH allows us to rapidly sequence whole genomes by using a branch-and-bound approach that keeps complexity to a low level. Not only is this approach fast and cheap but it also is very accurate. A rate as low as one mistake per million base pairs can be expected. Most importantly, thanks to the use of optical mapping, it is now possible to get haplotypes. There is still room for improvement in the SMASH program. For one thing, mistakes in the assembly will occur when the underlying search tree needs to be pruned. We could perform a second run where we focus on those locations where the tree had to be pruned and allocate more resources as to further expand the tree to be sure we get the proper path. There is also a nice theoretical analysis that can be done on the design of probes to justify what we have seen in the simulations.

But let us not lose our focus. Sequencing the whole genome is the cornerstone of any population study but it provides only the basis for these studies. Once the sequences are obtained, we need to do something with them. Some very important questions deserve to be asked. How important are haplotypes? Does it suffice to impute the haplotype-phasing from a population? How much information is captured by the known genetic variants (e.g., SNPs and CNVs)? How does one find the de novo mutations and their effects on various complex traits? Can exon-sequencing be sufficiently informative?

The other half of this dissertation has discussed the current state of knowledge on these questions. To develop personalized medicine, determining whether common, rare or a combination of both types of variants are responsible for common diseases seems to be a major step. I have developed a population genetics model that will be able to test different disease models. This model and its usage is still at an embryonic stage and needs some developing but the bases are solid and it will be easy to take over and keep improving it. The model allows one to simulate any population size evolution and any kind of disease. One obvious improvement would be to create non-random mating patterns such as an island model. It would also be interesting to study linkage relationships between SNPs under varying conditions of linkage disequilibrium.There is still a lot to do there but there is potential for a rewarding result.

As stated earlier, we are at a cross-roads. We may end up having to admit that the individualized analysis of sequences will not be able to bring us any useful information. But let us not forget that sequencing technologies, the very core of any further discovery in genetics, have been developing quickly and trying to optimize different constraints of the problem (accuracy, cost, rapidity, etc). From Sanger to nanopore technologies, many creative and innovative technologies such as pyrosequencing, sequencing by ligation, sequencing by synthesis have seen light. Unfortunately, none of these technologies have provided conclusive, error free sequencing results. I believe that the technology developed in this thesis will bring new life to the field and will give hope back to many physicians and

biologists. Furthermore, the ability to simulate different disease models may lead to a better understanding of how diseases work, in order to plan and evaluate results when real population studies are conducted.

Appendix A

Branch and Bound Efficiency

We have seen in the results that the algorithm works beyond what most people have expected in term of accuracy. An error rate of 1 base for every 10 000 base pairs is generally an acceptable rate for most studies and we can actually achieve an error rate of 1 per million base pairs. The problem is NP-complete and yet, we have extremely good efficiency. The underlying idea behind our technology is to create easy-to-solve instances of the PSBH problem. As stated above, this problem can be solved in a polynomial time if the probes do not hybridize more than two times on the sequence. This is very unlikely for long sequences but not for a restriction fragment of our sequence. Using 6-cutters, the expected length of our fragments is around 4000 bp. Using 6-mers, the probability that every 6-mer appears more than two times within the restriction fragment is very low, and we can treat each restriction fragment independently of the others. We are now asked to solve the PSBH multiple times (as many times as there are restriction fragments) but each of those instances of the problem is easy to solve.

Once we are able to get those small fragments, we are actually performing an exhaustive search with a Bayesian scheme of all possible assemblies of these small fragments, leading us to be quite confident we will get the correct assembly at the end of the search. This approach is motivated by the fact that we want to give each solution a chance. The counterpart of this is that we have to be sure our

tree does not grow exponentially. Assuming a random sequence, we can analyze the branching factor of our algorithm. Every node of our tree is extended by any of the four possible bases, given a probe that can be located within $\pm K$bp of the current location. A probe that occurs every Pbp (for 6-mers, P averages 4096) can be located every $\frac{P}{2}$ base pairs, including bases in the reverse complement. For each possible extension of the tree, the probability of finding a particular probe within our window of acceptance is therefore $\frac{4K}{P}$. Since each node is extended by four possible bases, the expected branching factor of a node is $\frac{16K}{P}$. If we want the number of branches generated to remain bounded, we have to keep this branching factor below 1.

Along the correct path, each node will have one correct extension and $\frac{12K}{P}$ random ones. Hence, the expected number of surviving branches will be $1 + \frac{\frac{12K}{P}}{1-\frac{16K}{P}}$. For example, if $K = 200$ and $P = 4096$, $\frac{16K}{P} = 0.781$, the expected number of surviving branches will be 3.68 which is a reasonable number. However, if $K = 250$, then the expected number of branches will be 32.24 (the expecting branching factor will be 0.976) and for $K = 255$, the number of branches becomes unbounded. This sudden jump from reasonable to undoable forces us to carefully choose the size of our window.

Appendix B

SMASH-P is NP-complete

Historically, sequencing by hybridization has been linked with graph theory problems, in particular finding an Eulerian path within a de Bruijn graph. The problem with a sequencing by hybridization sequencer and assembler was the non-uniqueness and ambiguity of the answer. The hope with positional sequencing by hybridization was that the extra information about the location of the probes would decrease this ambiguity. Unfortunately, we can prove that if the probes have more than 2 possible locations, the problem becomes NP-complete. Because there is a strong relationship between SBH and finding a Eulerian path in a graph, we will reduce the Positional Sequencing by Hybridization (PSBH) problem, described in [**?**] problem to the Positional Eulerian Path (PEP) problem. First, let us show that PEP is NP-complete. It will then be straightforward to reduce the PSBH problem to the PEP problem.

The PEP problem is to find an Eulerian path in a graph in which the edges of the path have to follow a certain order. Every edge e in a graph G is labelled with an integer L_e which represents the location of the probe. A positional Eulerian path is a path in which the position of the edge e, P_e, matches L_e. We can relax this assumption a little bit and allow P_e to be within a window of size W relative to L_e. Mathematically speaking, $|P_e - L_e| \leq W$. To prove this problem to be NP-complete, we can reduce it to the well known Hamiltonian path problem in a directed graph.

Let us start with a graph $G(V, E)$ such that the in-degree and the out-degree are equal to 2 for every vertex. Therefore, with $|V| = n$ we have $|E| = 2n$. Let us fix $W = 4n$. We build a graph $G'(V', E')$ with $|V'| = 4|V|$ and $|E'| = 3|E|$ as follows:

- We split every vertex u_i of G into three vertices $(u_{i,1}, u_{i,2}, u_{i,3})$.
- Every $u_{i,1}$ has an edge directed to $u_{i,3}$ and $u_{i+1,1}$ (for the vertex $u_{n,1}$, the vertex $u_{n+1,1}$ is the vertex $u_{1,1}$ which will always be the case later on). There are $2n$ such edges and their location P_e is $6n$. Their window of accepted location is then $\{2n, 6n\}$.
- Every vertex $u_{i,3}$ has two edges directed to the vertex $u_{i+1,2}$. Those edges are the ones from the graph G, therefore, we have $2n$ such edges and their location P_e is $2n$. Their window of accepted location is then $\{2, 6n\}$.
- Finally, every vertex $u_{i,2}$ has an edge directed toward $u_{i,1}$ and $u_{i,3}$. That gives us our final $2n$ edges. The edges from $u_{i,2}$ to $u_{i,3}$ have location $P_e = 1$ and the ones from $u_{i,2}$ to $u_{i,1}$ have location $P_e = 6n$. Their windows of accepted location are then respectively $\{1, 2n\}$ and $\{2n, 6n\}$.

We will show that G has an Hamiltonian path $\Leftrightarrow$ G' has a positional Eulerian path.

$\Rightarrow$: Following the previous construction of G' from a graph G with a Hamiltonian path, here is how we construct a positional Eulerian path in G'. Starting at vertex $u_{1,2}$, we alternate edges from $u_{i,2}$ to $u_{i,3}$ and edges from $u_{i,3}$ to $u_{i+1,2}$ and we stop at $u_{n,3}$. Those edges are either labelled 1 or $2n$. The positional

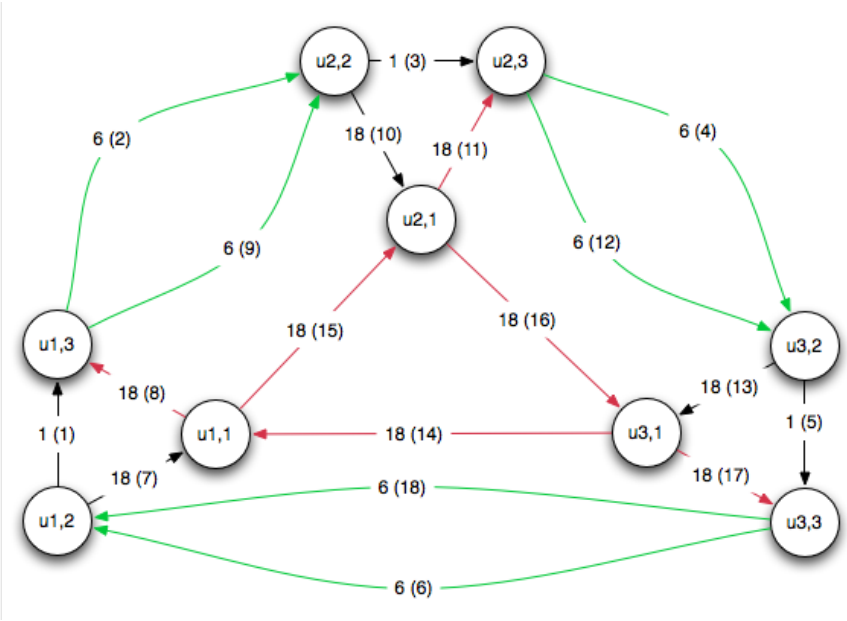

Figure 9: Example with a 3 vertices graph. Red edges: directed from $u_{i,1}$ to $u_{i,3}$ and $u_{i+1,1}$. Green edges: directed from $u_{i,3}$ to $u_{i+1,2}$. Black edges: directed from $u_{i,2}$ to $u_{i,1}$ and $u_{i,3}$. Numbers on the edges represent their location and numbers between parenthesis represent their position in the Eulerian path.

constraint ($|P_e - L_e| \leq W$) is respected since we start with an edge labelled 1 and we visit $2n - 1$ edges. Now, if we remove those edges from the graph, the remaining graph will be connected and every vertex will have equal in-degree and out-degree, except for the starting and the ending vertices which does not create a problem, and hence has an Eulerian path. The window of accepted location of the remaining edges provides that the Eulerian path in this remaining graph fits our positional assumption.

$\Leftarrow$: If G' has a positional Eulerian path, then construct a Hamiltonian path this way. For every $u_{i,j}$ vertices, go from vertex $u_{i,2}$ to $u_{i,1}$ and from $u_{i,1}$ to $u_{i,3}$. Then, go from $u_{i,3}$ to $u_{i+1,2}$ and repeat. End at $u_{n,3}$. You will have visited every node

once and only once.

Appendix C

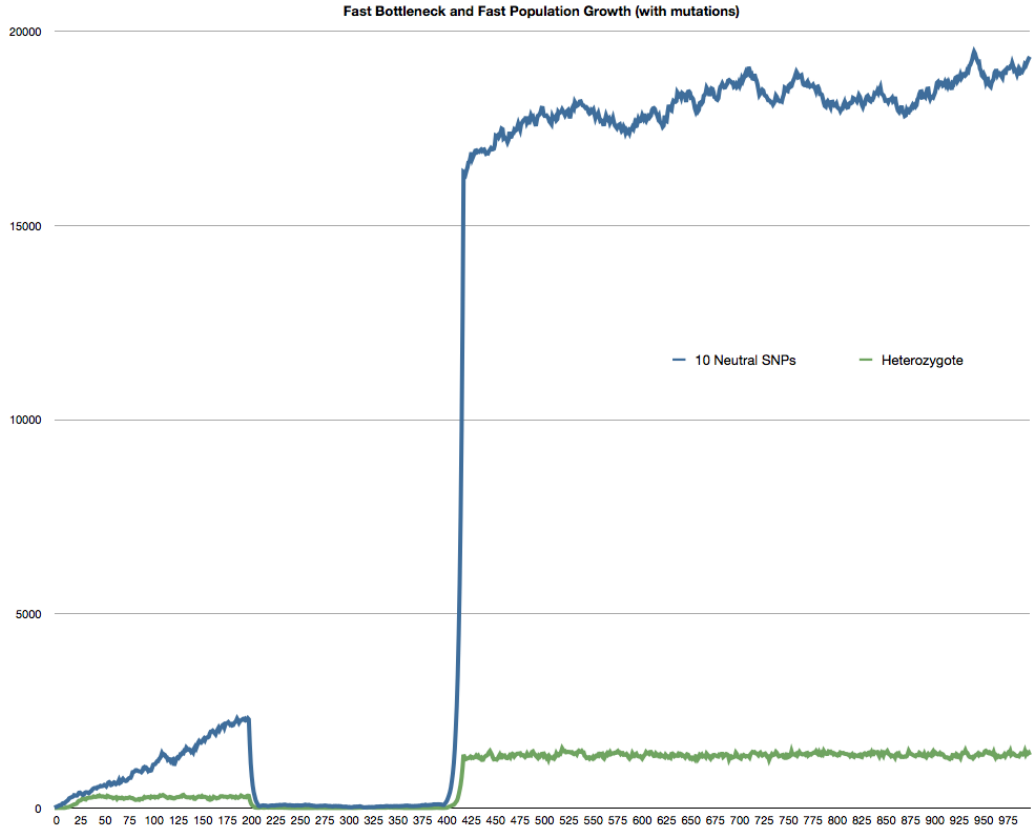

Figure 10: Here, we follow 10 SNPs that have an implication in a common disease. As long as one does not carry more than three of those SNPs, the individual will survive. If he has more than 3, he will die and not give birth to any offspring. We also follow a SNP known to give a heterozygote advantage to the carrier. The blue curve represents the total number of those 10 SNPs within the population while the green curve is the number of homozygote individuals. The population follows an abrupt bottleneck after 200 generations, leading the population from a 1000 individuals to as few as 27 in just 10 generations. The population remains constant for the next 190 generations before a rapid population expansion occurs. In 20 generations, the population count grows from 27 to 4888 individuals. As we can see in the figure, even if new mutations occur every new generation, the total number of SNPs or heterozygote individual reaches an equilibrium.

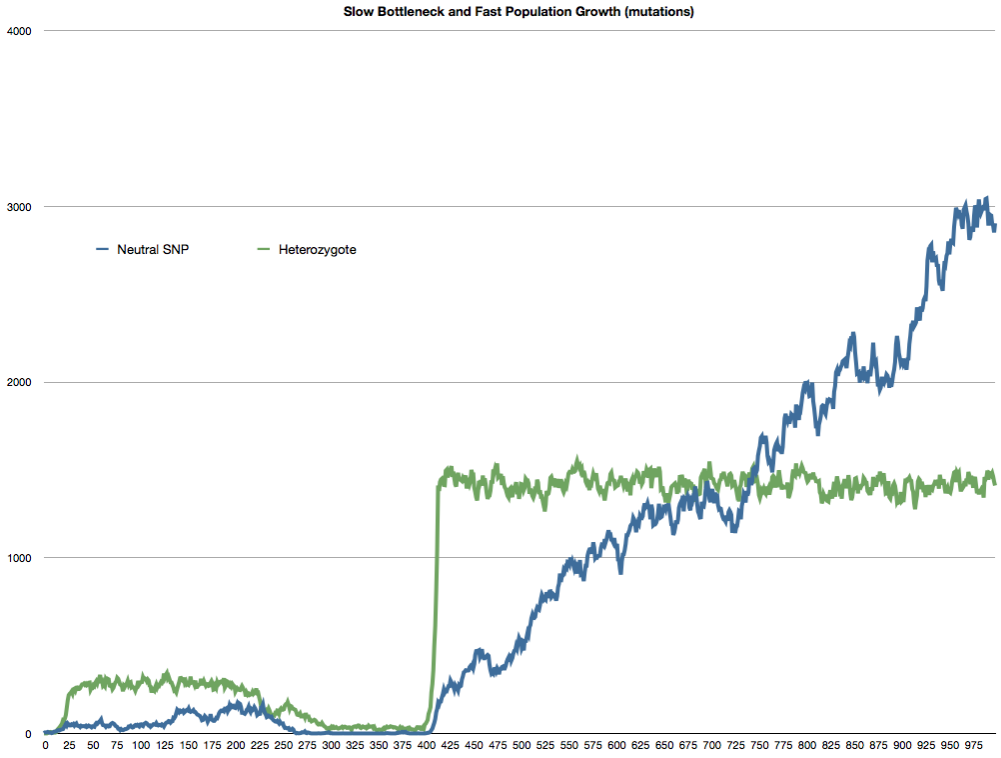

Figure 11: Here, we follow 1 SNP that has an implication in a common disease. We also follow a SNP known to give a heterozygote advantage to the carrier.The blue curve represents the total number of the followed SNP within the population while the green curve is the number of homozygote individuals. The population follows a slow bottleneck after 200 generations, leading the population 1000 individuals to 10 in 100 generations. The population remains constant for the next 100 generations before a rapid expansion. In 25 generations, the population count grows from 10 to 5426 individuals. As we can see in the figure, even if new mutations occur every new generation, the number of heterozygote individual reaches an equilibrium. We also see that the curves follow the evolution of the population size (a slow decrease and a quick increase). The total number of SNPs in the population keep growing since no selection effect is acting. The blue curve would reach fixation eventually.

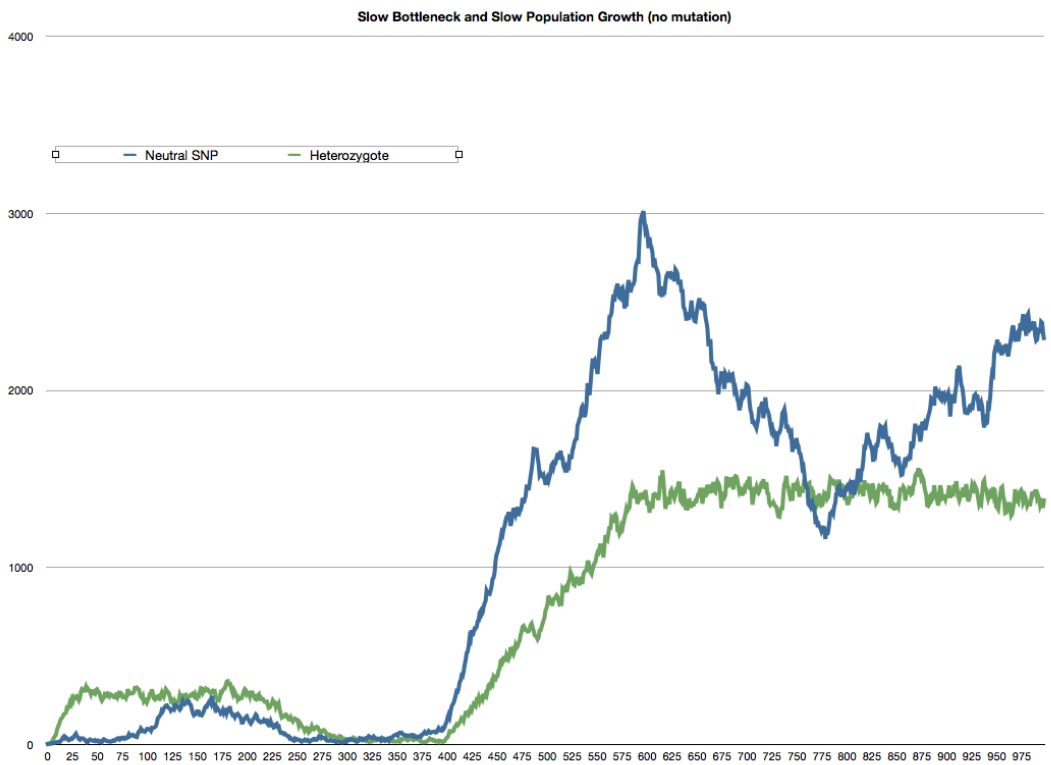

Figure 12: Here, we follow 1 SNP that has an implication in a common disease. We also follow a SNP known to give a heterozygote advantage to the carrier.The blue curve represents the total number of the followed SNP within the population while the green curve is the number of homozygote individuals. After 200 Generations, no new mutations are introduced in the population. The population follows a slow bottleneck after 200 generations, leading the population from a 1000 individuals to 10 in 100 generations. The population remains constant for the next 100 generations before a slow growth rate occurs. In 200 generations, the population count grows from 10 to 4010 individuals. Both curves follow the changes in population size. The number of heterozygote individuals still reaches an equilibrium. While the Hardy-Weinberg equilibrium states that the total number for the SNP followed should reach equilibrium, we see that it is not the case. This is probably due to a small size population combined with genetic drift and recombinations.

Bibliography

[1] J. Altmuller, L. J. Palmer, and G. Fischer et al. Genome wide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.*, 69:936–950, 2001.

[2] D. Altshuler, V. J. Pollara, and C. R. Cowles et al. A snp map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407:513–516, 2000.

[3] S. J. Chanock andT. A. Manolio and M. Boehnke et al. Nci-nhgri working group on replication in association studies. replicating genotype-phenotype associations. *Nature*, 447(7145):655–660, 2007.

[4] B. O. Bengtsson and G. Thomson. Measuring the strength of associations between hla antigens and diseases. *Tissue Antigens*, 18:356–363, 1981.

[5] J. Blangero. Localization and identification of human quantitative trait loci: king harvest has surely come. *Curr. Opin. Genet. Dev.*, 14:233–240, 2004.

[6] K. H. Buetow, M. N. Edmonson, and A. B. Cassidy. Reliable identification of large numbers of candidate snps from public est data. *Nat. Genet.*, 21:323–325, 1999.

[7] J. Butler, I. MacCallum, and M. Kleber et al. Allpaths: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18:810–820, 2008.

[8] L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nature Rev. Genet.*, 3:91–99, 2001.

[9] W. Casey, B. Mishra, and M. Wigler. Placing probes along the genome using pair-wise distance data. *Algorithms in Bioinformatics*, LNCS 2149:52–68, 2001.

[10] A. G. Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol. Biol. Evol.*, 7:111–122, 1990.

[11] F. S. Collins, A. Patrinos, and E. Jordan et al. New goals for the u.s. human genome project. *Science*, 282:682–689, 1998-2003.

[12] H. de Jong. Visualizing dna domains and sequences by microscopy: a fifty-year history of molecular cytogenetics. *Genome*, 46:943?946, 2003.

[13] V. Demidov. Pna and lna throw light on dna. *Trends in Biotechnology*, 21(1), January 2003.

[14] E. Eskin, E. Halperin, and R. M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinform. Comput. Biol.*, 1:1–20, 2003.

[15] A. Ben-Dor et al. On the complexity of positional sequencing by hybridization. *J. Comp. Bio*, 8(4):361–371, Jan 2001.

[16] A Lim et al. Shotgun optical maps of the whole escherichia coli o157:h7 genome. *Genome Research*, 11(9):1584–93, Sep 2001.

[17] A. Simeonov et al. Single nucleotide polymorphism genotyping using short, fluorescently labeled locked nucleic acid (lna) probes and fluorescence polarization detection. *Nucleic Acids Research*, 30(17):e91, 2002.

[18] B. Kerem et al. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245:1073–1080, 1989.

[19] C. Cantor et al. Sbh: an idea whose time has come. *Genomics*, 11, 1992.

[20] C. S. Carlson et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74:106–120, 2004.

[21] D. G. Wang et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280:1077–1082, 1988.

[22] D. Levy et al. Evidence for gene influencing blood pressure on chromosome 17. genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension*, 36:477–483, 2000.

[23] E. Halperin et al. Handling long targets and errors in sequencing by hybridization. *J. Comp. Bio.*, 10(3?4):483–497, 2003.

[24] F. Preparata et al. On the power of universal bases in sequencing by hybridization. *Proceedings of CIBM*, 3:295–301, 1999.

[25] F. Preparata et al. Sequencing-by-hybridization at the information-theory-bound: An optimal algorithm. *Brown University Tech. report,*, 1999.

[26] H. Stefansson et al. Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.*, 71:877–892, 2002.

[27] I. Smolina et al. Sequence-universal recognition of duplex dna by oligonucleotides via pseudocomplementarity and helix invasion. *Chemistry & Biology*, 10:591–595, July 2003.

[28] J. F. Gusella et al. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 306:234–238, 1983.

[29] J. Jing et al. Automated high resolution optical mapping using arrayed, fluid fixated, dna molecules. *Proc. Natl. Acad. Sci. USA*, 95:8046–8051, 1998.

[30] J. Lin et al. Whole-genome shotgun optical mapping of deinococcus radiodurans. *Science*, 285:1558–1562, Sept 1999.

[31] J. P. Hugot et al. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*, 411:599–603, 2001.

[32] L. Nistico et al. The ctla-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.*, 5:1075–1080, 1996.

[33] M. Koenig et al. Complete cloning of the duchenne muscular dystrophy (dmd). cdna and preliminary genomic organization of the dmd gene in normal and affected individuals. *Cell*, 50:509–517, 1987.

[34] M. Koizumi et al. Triplex formation with 2?-),4?-c-ethylene-bridged nucleic acids (ena) having c3?-endo conformation at physiological ph. *Nucleic Acids Research*, 31(12):3267–3273, 2003.

[35] N. J. Cox et al. Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am. J. Hum. Genet.*, 69:820–830, 2001.

[36] P. I De Bakker et al. Efficiency and power in genetic association studies. *Nature Genet.*, 37:1217–1223, 2005.

[37] R. J. Lipshutz et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechnics*, 19:442–447, 1995.

[38] S. Batzoglu et al. Arachne: a whole-genome shotgun assembler. *Genome Res.*, 12:177–189, 2002.

[39] S. John et al. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.*, 75:54–64, 2004.

[40] S. Zhou et al. A whole-genome shotgun optical map of yersinia pestis strain kim. *Appl. Environ. Microbiol.*, 68(12):6321–6331, 2002.

[41] T. Anantharaman et al. Genomics via optical mapping iii: Contiging genomic dna and variations. *ISMB99*, Aug 1999.

[42] T. Anantharaman et al. A probabilistic analysis of false positives in optical map alignment and validation. *WABI2001*, Aug 2001.

[43] T. J. Albert et al. Light-directed $5' \rightarrow 3'$ synthesis of complex oligonucleotide microarrays. *Nucleic Acids Res.*, 31:e35, 2003.

[44] X. Huang et al. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14, 1992.

[45] X. Huang et al. An improved sequence assembly program. *Genomics*, 33, 1996.

[46] Z. Lai et al. A shotgun sequence-ready optical map of the whole plasmodium falciparum genome. *Nature Genetics*, 23(3):309–313, 1999.

[47] D. M. Evans and L. R. Cardon. Guidelines for genotyping in genome wide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am. J. Hum. Genet.*, 75:687–692, 2004.

[48] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112, 1972.

[49] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, 12:921–927, 1995.

[50] R. A. Fisher. The genetical theory of natural selection. *Oxford*, 1930.

[51] K. A. Frazer, D. G. Ballinger, and D. R Cox et al. International hapmap consortium. a second generation human haplotype map of over3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[52] D. F. Gudbjartsson, D. O Arnar, and A. Helgadottir et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*, 448(7151):353–357, 2007.

[53] H. Harris. Enzyme polymorphism in man. *prsb*, 164:298–310, 1966.

[54] P. W. Hedrick. Genetic disequilibrium measures: proceed with caution. *Genetics*, 117:331–341, 1987.

[55] W. G. Hill. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 33:229–239, 1974.

[56] R. Idury and M. S. Waterman. A new algorithm for dna sequence assembly. *Journal of Computational Biology*, 2:291–306, 1995.

[57] G. Jimenez-Sanchez, B. Childs, and D. Valle. Human disease genes. *Nature*, 409:853–855, 2001.

[58] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.

[59] J. F. C. Kingman. Exchangeability and the evolution of large populations. *Proceedings of the International Conference on Exchangeability in Probability and Statistics*, pages 97–112, 1982.

[60] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19A:27–43, 2982.

[61] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*, 22:139–144, 1999.

[62] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genet.*, 27:234–236, 2001.

[63] E. S. Lander. The new genomics: global views of biology. *Science*, 274:536–539, 1996.

[64] Jeffrey M. Levsky and Robert H. Singer. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, 116 (14):2833–, 2003.

[65] R. C. Lewontin. The interaction of selection and linkage.i. general considerations; heterotic models. *Genetics*, 49:49–67, 1964.

[66] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations of *Drosophilia pseudoobscura*. *J. Comp. Bio*, 8(4):361–371, Jan 2001.

[67] R. C. Lewontin and K. Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:458–472, 1960.

[68] R. Li, H. Zhu, and J. Ruan et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, 20:265–272, 2009.

[69] K. E. Lohmueller, C. L. Pearce, M. Pike, and E. S. Lander. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common diseases. *Nature Genet.*, 33:177–182, 2003.

[70] G. Malécot. Mendélisme et consanguinité. *C. R. Acad. Sci.*, 215:313–314, 1942.

[71] T. A. Manolio, J. E. Bailey-Wilson, and F. S. Collins. Genes, environment and the value of prospective cohort studies. *Nat. Rev. Genet.*, 7(10):812–820, 2006.

[72] S. A. McCarroll and D. M. Altshuler. Copy-number variation and association studies of human disease. *Nat. Genet.*, 38(7):S37–S42, 2007.

[73] B. Mishra. Comparing genomes. *Special issue on Biocomputation: Computing in Science and Engineering*, pages 42–49, January/February 2002.

[74] P.A.P Moran. Random process in genetics. *Proc. Camb. Phil. Soc.*, 54:60–71, 1958.

[75] A. P. Morris. A flexible bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am. J. Hum. Genet.*, 79:679–694, 2006.

[76] G. Narzisi. Scoring-and-unfolding trimmed tree assembler: Algorithms for assembling genome sequences accurately and efficiently. 2011.

[77] B. Padhukasahasram, P. Marjoram, and J. D. Wall et al. Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, 178:2417–2427, 2008.

[78] P. A. Pevzner and R. J. Lipshutz. Towards dna sequencing chips. *19th Symposium on Mathematical Foundation in Computer Science*, 841, 1994.

[79] L. Picoult-Newberg, T. E. Ideker, and M. G. Pohl et al. Mining snps from est databases. *Genome Res*, 9:167–174, 1999.

[80] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant... or not? *Hum. Mol. Genet.*, 11:2417–2423, 2002.

[81] D. E. Reich, S. B. Gabriel, and D. Altshuler. Quality and completeness of snp databases. *Nat. Genet.*, 33:457–458, 2003.

[82] E. M. Reiman, J. A. Webster, and A. J. Myers et al. Gab2 alleles modifiy alzheimer's risk in apoe $\epsilon 4$ carriers. *Neuron.*, 54(5):713–720, 2001.

[83] M. J. Rieder, S. L. Taylor, and A. G. Clark et al. Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.*, 22:59–62, 1999.

[84] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517–820, 1996.

[85] R. Sachidanandam, D. Wesisman, and S. C. Schmidt et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933, 2001.

[86] J. T. Simpson, K. Wong, and S. D. Jackman et al. Abyss: A parallel assembler for short read sequence data. *Genome Research*, 19:1117–1123, 2009.

[87] J. C. Stephens, J. A. Schneider, and D. A. Tanguay et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, 2001.

[88] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68:978–989, 2001.

[89] G. Sutton. Tigr assembler: a new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1:9–19, 1995.

[90] H. K. Tabor, N. J. Risch, and R. M. Myers. Candidate-gene approaches for studying complex genetic traits. *Nature Rev. Genet.*, 3:391–397, 2002.

[91] P. Taillon-Miller, Z. Gu, and Q. Li et al. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res*, 8:748–754, 1998.

[92] S. Wacholder, S. Chanock, and M. Garcia-Closas et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Nat. Cancer Inst.*, 96:434–442, 2004.

[93] J. Wakefield. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.*, 81:208–227, 2007.

[94] J. D. Wall and J. K Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.*, 4:587–597, 2003.

[95] W. Y. S. Wang, B. J. Barratt, and D. G. Clayton et al. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev, Genet.*, 6:109–118, 2005.

[96] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *tpb*, 7:256–276, 1975.

[97] M. N. Weedon, G. Lettre, and R. M. Freathy et al. A common variant of hgma2 is associated with adult and childhood height in the general population. *Nat. Genet.*, 39(10):1245–1250, 2007.

[98] H. Weier. Dna fiber mapping techniques for the assembly of high-resolution physical maps. *The Journal of Histochemistry & Cytochemistry*, 49(8):939?948, 2001.

[99] J. West, J. Healy, M. Wigler, W. Casey, and B. Mishra. Validation of s. pombe sequence assembly by micro-array hybridization. *Journal of Computational Biology*, 13(1):1–20, Jan 2006.

[100] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[101] Z. Yang, G. K. Wong, and M. A. Eberle et al. Sampling snps. *Nat. Genet.*, 26:13–14, 2000.

[102] X. Zhu, C. A. McKenzie, and T. Forrester et al. Localization of a small genomic region associated with elevated ace. *Am. J. Hum. Genet.*, 67:1144–1153, 2000.