EXPERIMENTAL AND COMPUTATIONAL ADVANCES FOR STUDYING THE HUMAN GENOME WITH OPTICAL MAPPING

by

Brian P. Teague

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Cellular and Molecular Biology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2012

Date of final oral examination: September 20th, 2012

The dissertation is approved by the following members of the Final Oral Committee:

David C. Schwartz, Professor, Chemistry and Genetics
Michael N. Gould, Professor, Oncology
Michael Newton, Professor, Statistics and Biostatistics & Medical Informatics
Audrey Gasch, Associate Professor, Genetics
Patrick Krysan, Associate Professor, Horticulture

To my mother. I hope you remember my "defense" as fondly as I remember yours!

ACKNOWLEDGMENTS

I am indebted to my advisor for his guidance in my development as a scientist; to my labmates, from whom I have learned as much as I ever did in a classroom; and to my friends and family, without whose support and encouragement I would never have gotten this far.

CONTENTS

Contents iii

List of Tables v

List of Figures vi

Abstract ix

1 Background 1

1.1 A Brief History of Medical Genetics 2

1.2 The Genomic Era (Modern Day) 7

1.3 Thesis Overview 17

1.4 Bibliography 18

2 Optimization of Optical Mapping Protocols 39

2.1 Improved Methods for Optical Mapping 40

2.2 Optical Mapping of Three Normal Human Genomes 66

2.3 Conclusion 71

2.4 Bibliography 72

3 Structural Variation in Four Normal Human Genomes 81

3.1 Iterative Consensus Map Assembly 81

3.2 Four Thousand Structural Variants from Four Genomes 85

3.3 Conclusion 101

3.4 Bibliography 104

4 A Hidden Markov Model for Optical Map Analysis 109

4.1 A Hidden Markov Model for Optical Mapping 113

4.2 Common HMM Algorithms Validate the Optical Mapping HMM 121

4.3 Reference Map Refinement via *HMM Model Selection* 126

4.4 Haplotype Recovery with Belief Propagation 136

4.5 Conclusion 146

4.6 Bibliography 149

5 Conclusion 157

5.1 The Future of Personalized Medicine: From Risk Factors to Tailored Treatments 157

5.2 Cancer: The High Ground 160

5.3 Optical Mapping and its Successors 161

5.4 Conclusion 162

5.5 Bibliography 163

Colophon 175

LIST OF TABLES

1.1 Comparison of several platforms' abilities to detect structural variation . 13

2.1 The top 15 most diverse HapMap samples 67

2.2 Optical Mapping data collection statistics 71

3.1 Optical Mapping error processes 84

3.2 Optical map collection and assembly statistics 85

3.3 Summary of structural variants discerned by Optical Mapping 88

3.4 Summary of Optical Mapping results compared to other platforms 92

4.1 Performance of iterative assembly and HMM-based refinement compared 135

4.2 The belief propagation procedure uncovers loss of heterozygosity as embryonic stem cell lines are passaged 146

LIST OF FIGURES

2.1 An overview of the Optical Mapping platform 41
2.2 Basics of Optical Mapping 42
2.3 Lysis conditions effect DNA length, purity and homogeneity 46
2.4 Hydrolytic deposition of silanes 49
2.5 Drying conditions affect Optical Mapping surface properties 51
2.6 A microfluidic device for pump-mediated molecule deposition 54
2.7 Molecule stretch compared between pump-mounted and capillary-mounted DNA molecules 56
2.8 An example output of the collection dashboard 57
2.9 Acrylamide polymerization over time 59
2.10 Dashboard view of staining parameters 63
2.11 Variation in staining decreases with time 64

3.1 An overview of the map assembly pipeline 83
3.2 Iterative assembly extends and refines consensus maps 84
3.3 The Optical Map is highly accurate 86
3.4 Structural variation found in four genomes 89
3.5 A comparison of Optical Mapping results with end mapping methods . . 91
3.6 Optical Mapping's results show strong concordance with those of fosmid-end sequencing 93
3.7 One of the outliers discarded from the analysis in Figure 3.6 94
3.8 Optical Mapping finds large insertions that fosmid-end mapping misses . 95

3.9 Optical Mapping's results show a strong concordance with those of paired-end mapping 97
3.10 The optical map complements hybridization-based approaches. 99
3.11 Non-SNP extra cuts and missing cuts are likely small indels 100
3.12 Optical Mapping reveals variants inaccessible to other platforms 102

4.1 A basic Markov chain 110
4.2 A basic hidden Markov model 111
4.3 A hidden Markov model for Optical Mapping data analysis 117
4.4 Different HMM states represent different Optical Mapping errors 118
4.5 A single-molecule restriction map is represented by the path through the model that generated it. 120
4.6 The HMM discriminates between correct and incorrect models 123
4.7 The HMM can be used for pairwise alignment 124
4.8 The HMM can estimated unknown error model parameters 127
4.9 Using a hidden Markov model to refine a consensus restriction map. . . . 129
4.10 Using synthetic data to validate the map refinement algorithm. 132
4.11 The HMM-based refinement procedure correctly recovers changes to the reference 133
4.12 The HMM-based refinement procedure correctly estimates changed fragment sizes 134
4.13 For many variants, iterative assembly and HMM-based refinement perform similarly 135
4.14 Many of the variants found by one method but not the other appear to be heterzygous 136

4.15 An example factor graph for belief propagation 139
4.16 Belief propagation maximizes the factor graph's objective function. . . . 141
4.17 The factor graph assigning single-molecule haplotypes to Optical Mapping data. 142
4.18 Using synthetic data to validate the halotype recovery stragegy. 143
4.19 The belief propagation procedure correctly recovers synthetic haploid variation. 144
4.20 The belief propagation procedure recovers haploid variation from a human Optical Mapping data set 145
4.21 The belief propagation procedure uncovers loss of heterozygosity as embryonic stem cell lines are passaged 147

ABSTRACT

Many human geneticists were surprised by the discovery in 2004 that normal human genomes differ in structure as well as in sequence. Comprising up to 5% of the genome, these submicroscopic insertions, deletions, inversions and rearrangements represent a substantial source of genetic polymorphism and have been implicated in human evolution and disease. However, due to their large size and frequent association with repeated sequence, they remain poorly characterized by current methods.

This thesis addresses the discovery and characterization of genome structure variation in normal human genomes using Optical Mapping. Optical Mapping is a unique platform for analyzing genomes: it uses measurements of single molecules of DNA to infer a high-resolution genome-wide restriction map, whose representation of genome structure complements genome sequence to yield biological insight. These restriction maps are useful in a variety of genome analyses, including aiding in sequence assembly, probing cancer genomes for new therapeutic targets, and understanding normal human genetic variation.

The thesis begins with a careful optimization of many Optical Mapping protocols, with an eye towards improving throughput, consistency and data quality. Then, it describes the creation of genome-wide restriction maps of four normal human genomes, allowing us to analyze the structure of these genomes in unprecedented breadth and detail. The approach is validated by showing strong concordance with existing methods, while describing thousands of new variants with sizes ranging from kilobases to megabases of affected sequence.

The thesis concludes with the development of new analyses for Optical Mapping data sets based on a hidden Markov model (HMM). HMMs have found use in a variety

of bioinformatics endeavors including gene finding, copy number analysis, secondary structure prediction and multiple sequence alignment. The best-studied problems on hidden Markov models (evaluation, decoding, and learning) translate directly to common tasks in analyzing Optical Mapping data and provide a jumping-off point for addressing more interesting problems including restriction map refinement and haplotype discernment.

1 BACKGROUND

Modern medical genetics is founded on the centuries-old observation that an organism's traits are heritable. The family history has long been a staple of risk assessment in primary health care [1], reflecting the recognition of a genetic component in common diseases such as cardiovascular disease [2], Alzheimer's [3], diabetes [4], and numerous types of cancer [5–7]. The influence of genetic makeup on treatment efficacy has also been long recognized [8], though its predictive use is frequently held hostage to crude proxies such as racial heritage (e.g. [9]).

The completion of the Human Genome Project's draft sequence in 2001 [10, 11] heralded the first steps toward elucidating the connection between human genotype and phenotype on a genome-wide scale. It spawned a raft of technologies for collecting biological information in a global, hypothesis-free fashion: ways to profile *all* a tissue's RNA, for example, or *all* a cell's proteins. These "'omics" methods have granted researchers unprecedented power to measure both an individual's genetic makeup and its phenotypic consequence; however, a comprehensive mapping between the two remains to be established.

Work towards this goal depends crucially on a complete catalog of human genetic variation, yet recent discoveries suggest that our grasp of normal (i.e. non-pathogenic) human polymorphism remains incomplete [12, 13]. Over the last decade, the field of human genetics was blindsided by the revelation that normal genomes vary not only in *sequence*, as with single nucleotide polymorphisms, but also in *structure* [14–16]. These structural variants consist of multi-kilobase regions of sequence that are added, deleted, inverted or rearranged from one individual to the next [17, 18], affecting by some estimates upwards of 5% [19] of the euchromatic human genome. Following

hard on the heels of these variants' discovery came their association with human development [20, 21], physiology [22] and disease [23–25], highlighting their relevance to human health in particular and the study of genetics in general.

Despite its apparent importance, variation in human genome structure remains largely unexplored [13]. This thesis work represents an effort to address this gap in our understanding of human genomes and human genetics. At its heart lies Optical Mapping [26–35], a novel system for genome analysis based on high-throughput measurements of individual molecules of DNA. Described in detail later, Optical Mapping offers an unparalleled combination of genome-wide scope and sub-kilobase resolution for the analysis of genome structure. The advances in both experimental and computational methods described in this thesis better position Optical Mapping to shed light on this poorly understood class of human genetic variation.

1.1 A Brief History of Medical Genetics

Genetics as Natural History

Even though Gregor Mendel is rightly known as the father of modern genetics, the study of heredity and inherited disorders significantly predates his famous pea breeding experiments. As early as 1645, an English physician noted an apparently hereditary polydactyly [36](cited in [37]). A more definitive early account is that of Pierre Louis de Maupertuis, whose description of a familial polydactyly across four generations included a rudimentary statistical calculation demonstrating the improbability that the cluster of observed individuals was due solely to chance [38]. Perhaps the most famous example of a pre-Mendelian familial disorder is the eponymous chorea reported

in 1852 by George Huntington [39], who described not only its hereditary transmission but noted that there was no evidence of transmission by unaffected family members, the hallmark of an autosomal dominant gene.

Thus the contribution of Friar Mendel was not that traits are transmitted from one generation to the next: this had been recognized by breeders and physicians alike for centuries. Instead, his experiments demonstrated that genes were essentially atomic in nature, subject to the laws of segregation and independent assortment [40]. Though its importance went unrecognized during his lifetime, Mendel's work would receive wide attention at the turn of the century, due in large part to its promotion by the English zoologist William Bateson [37].

Mendel's theories would find their first clinical application in the work of Archibald Garrod, a London physician interested in the chemical basis of disease. Beginning in 1899, Garrod published a series of reports [41–43] on alkaptonuria, a rare but relatively harmless disease distinguished by darkened urine. In his 1901 paper, Garrod described several families in which alkaptonuria was highly consanguinous, families known to have resulted from the union of first cousins. In correspondence with Garrod, Bateson recognized this as precisely the scenario likely to give manifestation to recessive traits; together, the two had identified the first Mendelian disorder in humans.

Genetics as Biochemistry

Perhaps even more important than Garrod's discovery of the first Mendelian disease was his insight into biochemical basis for alkaptonuria and other metabolic disorders. (This was likely owing to his decade-long association Frederick Hopkins, who received

the Nobel Prize in Medicine for his work on metabolism and vitamins.) Over the next decade, Garrod developed the concept of "chemical individuality" and proposed that genetic diseases such as alkaptonuria represented "inborn errors of metabolism" [44], the biochemical counterparts of familial structural abnormalities such as polydactyly. His remarkably prescient monograph *The Inborn Factors in Disease* [45] went even further, suggesting that common "diatheses" (tendencies to suffer from common diseases) represented inherited differences in individual biochemistry and foreshadowing modern medical genetics' focus on genotype and disease risk.

It would be another three decades before Garrod's "inborn errors of metabolism" were reproduced in a tractable model system: Beadle and Tatum's work on *Neurospora*, using X-ray mutagenesis to create strains that lacked specific nutritional abilities [46]. At last, the genetic basis of these well-characterized biochemical pathways could be analyzed experimentally, giving new credence to the "one gene, one enzyme" hypothesis. Beadle and Tatum's work served to reconnect the study of biochemistry and metabolism with that of "classical" genetics (e.g. mapping, multiple alleleism, and population genetics), pioneered by Thomas Hunt Morgan's famous "Fly Room" at Columbia [47].

The "one gene, one enzyme" hypothesis was also the starting point for the application of classical genetics to human disease. In the decades since Garrod's identification of alkaptonuria as a Mendelian trait, steady progress had been made in identifying the biochemical bases of diseases such as phenylketonuria [48] and methemoglobinemia [49]. Bell and Haldane had also reported the first human genetic linkage, the X-linked association between hemophilia and color blindness [50]. However, efforts to create a comprehensive human gene map were severely hampered by the lack of reliable Mendelain polymorphisms with which to measure linkage. Researchers relied

primarily on biochemical markers measurable in blood, including blood groups [51], serum proteins [52], and HLA markers [53], in order to apply Morgan's work on linkage and mapping to human traits. Even after Watson and Crick's determination of DNA structure [54] provided an obvious molecular basis for inheritance, human genetics remained primarily biochemical – because without a method for amplifying small quantities of human DNA, it was infeasible to purify enough of it to study!

It comes as no surprise, then, that the first association of a gene's nucleotide sequence with its phenotypic consequence came in the context of hemoglobin. Pauling had identified sickle cell disease as a "molecular disease" of hemoglobin in 1949 [55], and in 1957 Ingram's peptide fingerprinting technique had discovered the amino acid change responsible [56]. With Temin and Baltimore's independent discoveries in 1970 of reverse transcriptase [57, 58], the abundant hemoglobin mRNA in peripheral blood reticulocytes could be readily converted to cDNA, which in turn could be hybridized to unamplified genomic DNA, isolating the hemoglobin gene for further study. In this manner, it was determined that α-thalassaemia was caused by the entire deletion of the α-globin genes [59, 60]. α-thalassaemia is also notable for being the disease diagnosed prenatally by molecular analysis [61], marking the first practical clinical application of human molecular genetics.

Genetics as Molecular Biology

The 1970s ushered in a host of new methods for studying nucleic acids. DNA cloning [62, 63], restriction digests [64] and Southern blotting [65] provided ways to amplify, manipulate and interrogate DNA sequences, giving geneticists the tools they needed to apply classical genetic techniques to Mendelian traits in humans. Key to this

development was the discovery by Botstein and colleagues of wide-spread restriction fragment length polymorphisms (RFLPs) [66] in human DNA. Relatively easy to genotype, RFLPs rapidly came to dominate the field of human genetics, displacing earlier biochemical markers. So pervasive and varied was this class of polymorphism that it promised enough information to compile a complete human genetic map.

RFLPs' prevalence also promised an immediate clinical application: an RFLP sufficiently close to a disease gene could serve as the basis for genetic prediction of that allele's genotype. A linked RFLP predictive of Duchenne muscular dystrophy was discovered in 1982 [67] and used soon thereafter in prenatal diagnosis [68]. RFLPs linked to cystic fibrosis [69] and Huntington's chorea [70] came soon after.

Finally, with the widespread characterization of DNA polymorphisms, it became possible not only to map disease-related genes but also to actually isolate them using positional cloning. A gene's sequence might be expected to provide insight into the molecular etiology of the disease; also, clinicians could look for relationships between individual sequence mutations and disease phenotypes (e.g. severity and age of onset), a process that continues today. Given the existence of a closely linked RFLP, it is no surprise that the gene behind Duchenne muscular dystrophy was the first to be cloned [71]; it was followed by the genes causing cystic fibrosis [72] and fragile X mental retardation [73, 74]. Positional cloning also resolved controversy over the mechanism behind "non-Mendelian" diseases such as Huntingtons [75], illuminating the molecular basis behind its classic "anticipation" phenotype [76].

By the late 1980s, enthusiasm over the possibility of a complete gene map had evolved into concrete plans to sequence the entire human genome. Initially a conception of the United States Department of Energy, the Human Genome Project rapidly grew to an international project with major support from the US National Institutes

of Health and the United Kingdom's Wellcome Trust. The completion of a draft sequence in 2000 [10, 11] and the report of the last "complete" chromosome sequence in 2006 [77] set the stage for modern genome-scale genetics.

1.2 The Genomic Era (Modern Day)

SNPs, GWAS and Common Disease

Despite the advances of the previous two decades, a century of traditional genetics had made it clear that most common diseases were non-Mendelian [78, 79] and not, in general, amenable to positional cloning [80]. There were a few rare successes, including the identification mutations that had sizable risk impact for breast cancer [81], hypertension [82] and diabetes [83]; but these advances turned on the identification of mendelian sub-phenotypes (such as early age of onset) and were not generalizable to the broader population [79, 84].

Thus it was with growing excitement that the field of medical genetics turned to single nucleotide polymorphisms (SNPs). Originally made accessible by the polymerase chain reaction [85–87] and already popular for locus-specific genotyping [88, 89], the widespread availability of high-throughput automated sequencing technology made SNPs a more enticing substrate for genome-scale human genetics than labor-intensive RFLP screening. SNPs also proved more abundant than RFLPs, with a common (minor allele frequency [MAF] $> 1\%$) SNP to be had on average once every 300 bases [90, 91]. Even more importantly, SNPs could be scored with massive parallelism; a single microarray experiment costing less than $1000 could determine an individual's genotype at hundreds of thousands of loci [92, 93].

How could wide-spread genotyping of SNPs be used to identify genes involved in common diseases? The most common approach, called a genome-wide association study (GWAS) [94, 95], depends crucially on two assumptions. The first assumption is what has come to be known as the *"common disease–common variant" (CD-CV) hypothesis* [79]. CD-CV postulates that Mendelian diseases are typically rare because of strong purifying selection: they have a negative impact on reproductive fitness. Common diseases such as hypertension and type II diabetes, on the other hand, have modest (if any) impact on reproductive fitness; thus, the alleles that contribute to susceptibility to these diseases are likely more common. So, to uncover an association with a common disease, it should be sufficient to genotype a large number of people for common alleles, rather than a metaphorical "needle in the haystack" search for many disparate uncommon alleles.

The second assumption relates to feasibility, and it centers on the phenomenon of *linkage disequilibrium:* because the rate of crossover is both relatively low [96] and not uniform across the genome [97], alleles that influence common diseases tend to stay linked to nearby loci for many generations. Thus, rather than genotype the full complement of 10 million common SNPs, it should be possible to measure a subset of carefully chosen tag SNPs, whose genotypes could then serve as proxies for nearby loci. The task of creating a genome-wide map of linkage disequilibrium using SNPs was undertaken by the International HapMap Project in 2003 [98]; their initial results, published two years later [99], served to inform the design of subsequent GWAS studies.

Thus, the stage was set for an explosion in GWAS studies of common traits and diseases. The last five years have seen a number of notable successes, including the identification of genes involved in diabetes [100], age-related macular degeneration

[101], Chron's disease [102], and even height [103]. In cases where Mendelian sub-phenotypes and positional cloning had already identified loci, these were reliably recapitulated, in addition to scads of additional genes providing new avenues for future work on these important public health issues.

On the other hand, there has been some concern voiced that GWAS experiments are not giving us the full picture. Critics [104, 105] generally note that the additive heritability explained by GWAS experiments are all less than 50%, and most are under 20% [106]. What is responsible for the missing heritability? There are several possibilities:

- First, it is possible that the CV-CD hypothesis is wrong [107]. In this case, disease alleles are either less common and more penetrant, leading to the need to genotype many more alleles to describe these common diseases' prevalance; or they are more common and less penetrant, and the association signal is getting lost in the noise of spurious low-level associations. It is important to note that these two possibilities are not mutually exclusive, either. As sequencing entire genomes replaces genotyping at preselected loci [108–112], it should become clear whether either (or both) of these possibilities is true.

- Because each GWAS experiment consists, at its core, of a very large number of statistical tests of association between allele and phenotype [113], a stringent multiple test correction procedure is required to filter out spurious associations. The most frequently employed procedure, the Bonferroni correction [114], assumes that each test is independent; given the linkage between nearby polymorphic loci, this assumed independence makes Bonferroni over-conservative. Thus, some of the "missing" heritability could be a result of sub-threshold

events that are being wrongly filtered out.

- Because the HapMap project was itself constrained both in the number of SNPs to genotype and the number of samples to test, it is possible that our understanding of inheritance and linkage is incomplete: a poor choice of tag SNPs might be undermining our ability to discern relevant associations [84]. The broader scope of the 1000 Genomes Project [112] should provide further insight into the extent and usability of linkage disequilibrium in large-scale studies.

- Finally, it is possible that some of the missing inheritance is simply illusory. The "additivity" of heritability is a poor proxy for epistasis, non-linear interactions between alleles, epigenetic contributions, interactions between genetics and the environment, etc.

Moreover, Lander argues persuasively [115] that the "hand wringing" over "missing" heritability in GWAS studies is likely misplaced, anyways. The results of a properly conducted GWAS study are the observational equivalent of a mutagenesis study in a model organism such as *Drosophila* or *Arabidopsis*: an unbiased, global, hypothesis-free way of asking which genes affect the process or disease the researcher is interested in. A prime example comes from the GWAS studies of variation in cholesterol [116, 117]: the *HMGCR* locus has a common variant at 40% frequency that explains only a modest 2.8 mg/dl average change in low-density lipoprotein (LDL) level. However, the HMG-CoA reductase encoded by *HMGCR* is the direct target of the statin drugs [118], a potent class of therapeutics for reducing LDL levels and the correspondent risk of myocardial infarction.

Structural Variation

The sequence of the entire human genome brought with it widely available technologies for investigating variation in genome structure as well as genome sequence. Early studies consisted of differentially labeled samples hybridized *in situ* to metaphase chromosomes [119], but resolution was on par with traditional karyotyping. Methods with improved resolution [120–123] were crucial since previous studies [59, 60] had shown that sub-microscopic changes in genome structure (i.e. too small to be visible on a karyotype) could be pathogenic. There was also substantial interest from the cancer community [121–124], because large-scale duplication and mosaicism is one of the hallmarks of a cancer genome.

Thus, it was with some surprise that investigators conducting a cohybridization experiment on a cancer sample in 2004 identified widespread copy number variation (CNVs) *in their normal controls as well* [15, 125]. The resolution of early hybridization-based methods was poor, but their tantalizing results (and the relative inexpensiveness of microarray experiments) drove a number of increasingly ambitious studies [126–129], culminating in a massive CNV discovery and genotyping effort by the Wellcome Trust Case Control Consortium [19].

At the same time, complementary approaches were being pioneered by our laboratory [130, 131] and by Eichler and colleagues at the University of Washington in Seattle. While our laboratory was focused on scaling up the Optical Mapping system to address human and similarly sized genomes (discussed in the next section), Eichler's group employed a large-scale clone-end mapping strategy [132, 133]: they created a large library of random clones, then sequenced their ends and "mapped" (aligned) those end sequences back to the reference genome sequence. If the clone

library is created in such a way that the clone sizes are tightly controlled (say, 40 kb or so), then any clone whose ends' alignments span significantly more or less sequence than the average clone size contains a structural variant. Clone-end mapping is several orders of magnitude more expensive than a microarray experiment, but has several advantages over hybridization-based approaches. Its resolution is better than all but the largest microarray experiments; it can discern *balanced* variants in which copy number does not change; and the variants are captured completely by the clone library, allowing for easy sequencing and nucleotide-resolution analysis.

Finally, with the wide-spread availability of next-generation sequencing, it was inevitable that sequencing methods would be applied to the hunt for human genome structural variation. Several entire human genomes have been sequenced using standard next-generation sequencing methods [108–111], and a number of other studies [134–137] have specifically interrogated genomes for structural variation with a paired-end mapping strategy similar to the clone-end mapping described above (though they sacrifice the ability to archive the clones in doing so.)

Unfortunately, it is not clear that the last 8 years have brought us to a complete understanding of structural polymorphism. Of greatest concern is the modest concordance between different studies using different technologies, even when applied to the same sample, and even when corrected for the technical limitations of the particular methods (Table 1.1). The lack of overlap is strongly suggestive of ascertainment bias, which follows naturally from consideration of the technologies being employed:

- Hybridization techniques are confounded by non-specific hybridization in repeat-rich regions. This insensitivity is particularly troublesome in relation to structural variants, which are frequently associated with repeats [138].

Query Platform	Reference Platform			
	Fosmid End-Sequencing	Paired-End Mapping	Affymetrix SNP 6.0	Tiling Array CGH
Fosmid End-Sequencing		92/196 (47%)	262/564 (46%)	262/564 (46%)
Paired-End Mapping	62/109 (57%)		146/163 (90%)	461/641 (72%)
Affymetrix SNP 6.0	562/9527 (6%)	173/753 (23%)		17628/217344 (8%)
Tiling Array CGH	686/9527 (7%)	631/826 (76%)	17628/217344 (8%)	

Table 1.1. Comparison of several platforms' abilities to detect structural variation A comparison of structural variant detection overlap between several technological platforms when applied to the same samples. Each cell shows the number of variants from the reference platform's results that were detected by the query platform. The reference platform's variants are first filtered to remove those that the query technology is not expected to be able to detect. Fosmid end-sequencing data from [132] and [133]; paired-end mapping data from [134]; Affymetrix CNV data from [128]; tiling array CGH data from [19].

- Clone-based strategies are limited by a maximum clone insert size and a wide clone insert size distribution relative to the size of the events they are trying to detect.

- Paired-end sequencing is hampered by both of the above problems: short read lengths limiginb usefulness in repeat-rich regions, and generally small insert sizes (just several kb).

What *is* clear from almost a decade's work on structural variation is that, while many variants appear benign, a significant number have been either associated with or identified as directly causing a number of diseases, including pancreatitis

[139], psoriasis [140], hypercholesterolemia [141] and susceptibility to lupus [142] and HIV-1/AIDS [23]. Interestingly, a preponderance of these associations have been neurological or neurodevelopmental in nature [25]: a microdeletion responsible for Prader-Willi syndrome was described in the 1980s [143], and has since been joined by both recurrent and non-recurrent copy number variants and rearrangements, including a recurrent microdeletion at 17p11.2-p12 causing developmental delay [144], a CNV at Xq28 associated with Rett's syndrome [145, 146], and a duplication at 21q21 that appears to enhance predisposition to Alzheimer's disease [147], among many others [25]. A number of recent studies have also described an association between structural variants and common complex neurocognitive disorders including autism [148], schizophrenia [149–151], epilepsy [152] and Parkinson disease [153, 154].

Finally, structural variants may have something to teach us about human evolution. Structural variants are strongly enriched in regions of the genome containing segmental duplications [138], which (via non-allelic homologous recombination) provides a likely mechanistic explanation for their occurrence [155, 156]. Such regions are also commonly show strong signatures of positive selection [157], and show functional enrichment for immune response, xenobiotic recognition and reproduction. Seeing as gene duplication is one of the primary mechanisms for the creation of new genes [158], these are tantalizing hints for a possible association between segmental duplications, structural variation and *Homo* evolutionary success [138, 157]. In fact, a recent study focusing on an inversion encompassing the *tau* gene showed signs of still being subject to positive selection in the modern-day population of Iceland [159].

Optical Mapping

This thesis describes the application of Optical Mapping to the problem of discovering and characterizing normal structural variation in the human genome. Optical Mapping is a platform that provides insight into genome *structure*, complementing that provided by genome *sequence*. Optical Mapping grew out of studies of the properties and dynamics of large DNA molecules, originally with an eye towards understanding pulse-field gel electrophoresis [160]. Its original incarnation [26] used fluid flow to stretch large DNA molecules dissolved in molten agarose, which were then fixed in place as the agarose gelled. Live microscopy of these DNA molecules being cleaved by restriction enzymes allowed for the laborious construction of physical maps based on these single-molecule observations.

A number of technological advances in the last two decades have dramatically boosted the resolution and throughput while retaining the same basic idea. Molten agarose for the presentation of DNA molecules was replaced first with polyacrylamide [161], then with silanized glass surfaces [28, 162, 163], obviating the necessity to view the enzymology as it occurred. Random deposition was replaced with ordered, directed flow first *via* controlled evaporation of a fluid droplet [28], then in a microfluidic device [31, 164]; this engendered reproducible molecule presentation and obviated manual microscopy, allowing its replacement by an automated, computer-controlled microscopy workstation [28, 31, 165–167]. Having the molecules all oriented in the same direction also enabled the replacement of laborious manual image markup by an automated machine-vision pipeline. These synergies, potentiated further still by sensitive CCD cameras, high quantum-yield intercalating dyes and laser illumination, have boosted the throughput of a modern Optical Mapping workstation to tens of

thousands of single molecule observations per day [35]. The single-molecule restriction maps thus produced serve as the grist for a suite of custom-developed software for Optical Map analysis and visualization [35, 168–178], completing a comprehensive system for genome structure analysis.

These increases in efficiency, in turn, have enabled the application of Optical Mapping to larger and larger projects. Initially, the system was used to create restriction maps of clones and individual yeast chromosomes [27, 168, 179, 180], very much in the footsteps of traditional multiple-digestion restriction mapping. These applications gave way to shotgun mapping of entire bacterial genomes [29, 30, 166, 167, 181–184], frequently with the goal of aiding an ongoing sequencing project. Finally, the confluence of microfluidic molecule deposition, large-scale automation and reliable machine vision enabled the application of Optical Mapping to non-clonal, genomic samples from multi-gigabase genomes including important crops like rice [33] and maize [34, 185–187]; mammalian genomes [177]; and human genomes, both phenotypically normal [35, 188] and cancer-derived.

Importantly, Optical Mapping supplants older technologies while complementing modern modes of genome analysis. For example, traditional restriction mapping of large clones (or entire bacterial genomes) is a laborious process involving the construction of clone libraries and time-consuming, clone-by-clone digestion and electrophoresis. The modern Optical Mapping platform completely obviates such efforts; such experiments are the work of an afternoon. A high-resolution restriction map of an organism's genome can anchor a nascent sequencing effort [33, 34], especially useful given the short read lengths produced by next-generation sequencing instruments. In other contexts, such as cancer genomics, the genome-spanning high resolution restriction maps produced by Optical Mapping can bring important context

and biological insight to structural events that are poorly characterized by other technologies, or even missed altogether.

1.3 Thesis Overview

The previous two sections provide the rationale for the project that is the subject of this thesis: using Optical Mapping to discover and characterize human genome structural variation. Its unique synthesis of sophisticated chemistry, microfluidics, enzymology, instrumentation and computation give it unrivaled power to provide insight into genome structure. The remainder of this thesis, then, consists of a detailed account of the project and its results.

Chapters 2 and 3 describe the application of the current Optical Mapping system to the genomes of four phenotypically normal humans. Chapter 2 describes the Optical Mapping system in detail, including new directions in controlled fluid flow for DNA molecule deposition and quality-related metrics for ongoing collection projects. It also describes the data collection results for three of the four humans (the fourth having been collected by another student, S. Reslewic [131]).

Chapter 3 describes the analysis of the data thus acquired. It includes a detailed description of our iterative genome assembly process, the variants that were discovered, and an in-depth comparison to the sets of variants discovered using alternate methods.

Chapters 4 describes the development of novel computational methods for Optical Mapping analysis. It presents a hidden Markov model that was developed for the analysis of Optical Mapping data, including a detailed contrast with previous methods and the validation experiments that give us confidence in its results. It also shows the application of the hidden Markov model analyses to a number of Optical Mapping

data sets and describes the biological insight derived thereby.

Chapter 5 concludes this thesis with a summary of the results, a discussion of their importance and a prospectus for future work.

1.4 Bibliography

[1] Rich, E. C., et al. Reconsidering the family history in primary care. *Journal of general internal medicine*, **19**(3):273–80, 2004.

[2] Barrett-Connor, E. and Khaw, K. Family history of heart attack as an independent predictor of death due to cardiovascular disease. *Circulation*, **69**(6):1065–1069, 1984.

[3] Heston, L., et al. Dementia of the Alzheimer type. Clinical genetics, natural history, and associated conditions. *Archives of General Psychiatry*, **38**(10):1085–90, 1981.

[4] Hamman, R. Genetic and environmental determinants of nonâŁłinsulinâŁłdependent diabetes mellitus (NIDDM). *Diabetes/metabolism reviews*, **8**(4):287–338, 1992.

[5] Cramer, D., et al. Determinants of ovarian cancer risk. I. Reproductive experiences and family history. *J Natl Cancer Inst*, **71**(4):711–6, 1983.

[6] Sattin, R., et al. Family history and the risk of breast cancer. *JAMA : the Journal of the American Medical Association*, **253**(13):1908–13, 1985.

[7] Mecklin, J. Frequency of hereditary colorectal carcinoma. *Gastroenterology*, **93**(5):1021–5, 1987.

[8] Motulsky, A. G. Drug reactions, enzymes, and biochemical genetics. *JAMA : The Journal of the American Medical Association*, **165**(7):835–7, 1957.

[9] Exner, D. V., et al. Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction. *The New England journal of medicine*, **344**(18):1351–7, 2001.

[10] Lander, E. S., et al. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921, 2001.

[11] Venter, J. C., et al. The sequence of the human genome. *Science (New York, NY)*, **291**(5507):1304–51, 2001.

[12] Mills, R. E., et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**(7332):59–65, 2011.

[13] Baker, M. Structural variation: the genome's hidden architecture. *Nature Methods*, **9**(2):133–137, 2012.

[14] Iafrate, A. J., et al. Detection of large-scale variation in the human genome. *Nature genetics*, **36**(9):949–51, 2004.

[15] Sebat, J., et al. Large-scale copy number polymorphism in the human genome. *Science (New York, NY)*, **305**(5683):525–8, 2004.

[16] Eichler, E. E. Widening the spectrum of human genetic variation. *Nature genetics*, **38**(1):9–11, 2006.

[17] Feuk, L., Carson, A. R., and Scherer, S. W. Structural variation in the human genome. *Nature reviews Genetics*, **7**(2):85–97, 2006.

[18] Sharp, A. J., Cheng, Z., and Eichler, E. E. Structural variation of the human genome. *Annual review of genomics and human genetics*, **7**:407–42, 2006.

[19] Conrad, D. F., et al. Origins and functional impact of copy number variation in the human genome. *Nature*, **464**(7289):704–12, 2010.

[20] Lee, S., Cheran, E., and Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics (Oxford, England)*, **24**(13):i59–67, 2008.

[21] Walsh, T., et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, NY)*, **320**(5875):539–43, 2008.

[22] Perry, G. H., et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, **39**(10):1256–60, 2007.

[23] Gonzalez, E., et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science (New York, NY)*, **307**(5714):1434–40, 2005.

[24] Sharp, A. J. Emerging themes and new challenges in defining the role of structural variation in human disease. *Human mutation*, **30**(2):135–44, 2009.

[25] Stankiewicz, P. and Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual review of medicine*, **61**:437–55, 2010.

[26] Schwartz, D. C., et al. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (New York, NY)*, **262**(5130):110–4, 1993.

[27] Cai, W., et al. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(11):5164–8, 1995.

[28] Jing, J., et al. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(14):8046–51, 1998.

[29] Lin, J., et al. Whole-genome shotgun optical mapping of Deinococcus radiodurans. *Science (New York, NY)*, **285**(5433):1558–62, 1999.

[30] Lim, A., et al. Shotgun optical maps of the whole Escherichia coli O157:H7 genome. *Genome research*, **11**(9):1584–93, 2001.

[31] Dimalanta, E. T., et al. A microfluidic system for large DNA molecule arrays. *Analytical chemistry*, **76**(18):5293–301, 2004.

[32] Zody, M. C., et al. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature*, **440**(7087):1045–9, 2006.

[33] Zhou, S., et al. Validation of rice genome sequence by optical mapping. *BMC genomics*, **8**:278, 2007.

[34] Zhou, S., et al. A single molecule scaffold for the maize genome. *PLoS genetics*, **5**(11):e1000,711, 2009.

[35] Teague, B., et al. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24):10,848–53, 2010.

[36] Digby, K. *Two Treatises: In the One of which, The Nature of Bodies, In the other, The Nature of Mans Soule, Is Looked Into: In Way Of Discovery Of The Immortality of Reasonable Soules.* John Williams, London, 1645.

[37] Harper, P. S. *A Short History of Medical Genetics.* Oxford University Press, Oxford, 2008.

[38] Maupertuis, P. *Venus Physique (The Earthly Venus).* Johnson Reprint Corporation, english tr edition, 1753.

[39] Huntington, G. On chorea. *Med Surg Reporter (Phila)*, **26**:320–1, 1872.

[40] Mendel, G. Versuche uber Pflanze-hybriden (Experiments on plant hybrids.). *Verhandlung des Natur-forscheden Vereines in Brunn (Proceedings of the Natural History Society of Brunn)*, **4**:3–47, 1866.

[41] Garrod, A. A contribution to the study of alkaptonuria. *Med-Chir Trans*, **82**:369–94, 1899.

[42] Garrod, A. About alkaptonuria. *Lancet*, **2**:1484–6, 1901.

[43] Garrod, A. The incidence of alkaptonuria: a study in chemical individuality. *Lancet*, **2**:1616–20, 1902.

[44] Garrod, A. *Inborn Errors of Metabolism.* Hodder & Stoughton, London, 1909.

[45] Garrod, A. *The Inborn Factors in Disease.* Clarendon Press, Oxford, 1931.

[46] Beadle, G. W., Tatum, E. L., and Control, G. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, **27**(11):499–506, 1941.

[47] Morgan, T. H. *The Theory of the Gene*. Yale University Press, New Haven, CT, 1926.

[48] Folling, A. Ueber Asscheidung von Phenylbrenztraubensaeure in den Harn als Stoffwechselanomalie in Verbindung mit Imbezillitaet. *Ztschr Physiol Chem*, (227):169–76, 1934.

[49] Gibson, Q. H. The reduction of methaemoglobin in red blood cells and studies on the cause of idiopathic methaemoglobinaemia. *The Biochemical journal*, **42**(1):13–23, 1948.

[50] Bell, J. and Haldane, J. B. The linkage between the genes for colour-blindness and haemophilia in man. *Proc R Soc Lond B*, **123**:119–150, 1937.

[51] Bernstein, F. Zussamenfassende Betrachtungen uber die erblichen blutstrukturen die Menschen. *Z Indukt Abstamm Vereblehre*, **37**:237–70, 1927.

[52] Smithies, O. and Walker, N. F. Genetic control of some serum proteins in normal humans. *Nature*, **176**(4496):1265–6, 1955.

[53] Dausset, J. Iso-leuco-anticorps. *Acta Haematol*, **20**:156–66, 1958.

[54] Watson, J. D. and Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356):737–8, 1953.

[55] Pauling, L., et al. Sickle cell anemia a molecular disease. *Science (New York, NY)*, **110**(2865):543–8, 1949.

[56] Ingram, V. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*, **180**(4581):326–8, 1957.

[57] Temin, H. M. and Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, **226**(5252):1211–3, 1970.

[58] Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, **226**(5252):1209–11, 1970.

[59] Ottolenghi, S., et al. The severe form of alpha thalassaemia is caused by a haemoglobin gene deletion. *Nature*, **251**(5474):389–92, 1974.

[60] Taylor, J. M., et al. Genetic lesion in homozygous alpha thalassaemia (hydrops fetalis). *Nature*, **251**(5474):392–3, 1974.

[61] Kan, Y. W., et al. Successful application of prenatal diagnosis in a pregnancy at risk for homozygous beta-thalassemia. *The New England journal of medicine*, **292**(21):1096–9, 1975.

[62] Jackson, D. A., Symons, R. H., and Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **69**(10):2904–9, 1972.

[63] Cohen, S. N., et al. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, **70**(11):3240–4, 1973.

[64] Nathans, D. and Smith, H. O. Restriction endonucleases in the analysis and restructuring of dna molecules. *Annual review of biochemistry*, **44**:273–93, 1975.

[65] Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of molecular biology*, **98**(3):503–17, 1975.

[66] Botstein, D., et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, **32**(3):314–31, 1980.

[67] Murray, J. M., et al. Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature*, **300**(5887):69–71, 1982.

[68] Bakker, E., et al. Prenatal diagnosis and carrier detection of Duchenne muscular dystrophy with closely linked RFLPs. *Lancet*, **1**(8430):655–8, 1985.

[69] Knowlton, R. G., et al. A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature*, **318**(6044):380–2, 1985.

[70] Gusella, J. F., et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, **306**(5940):234–8, 1983.

[71] Koenig, M., 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**(3):509–17, 1987.

[72] Rommens, J. M., et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science (New York, NY)*, **245**(4922):1059–65, 1989.

[73] Oberlé, I., et al. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science (New York, NY)*, **252**(5010):1097–102, 1991.

[74] Fu, Y. H., et al. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell*, **67**(6):1047–58, 1991.

[75] MacDonald, M. E., et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, **72**(6):971–83, 1993.

[76] Walker, F. O. Huntington's disease. *Lancet*, **369**(9557):218–28, 2007.

[77] Gregory, S. G., et al. The DNA sequence and biological annotation of human chromosome 1. *Nature*, **441**(7091):315–21, 2006.

[78] Weeks, D. E. and Lathrop, G. M. Polygenic disease: methods for mapping complex disease traits. *Trends in genetics : TIG*, **11**(12):513–9, 1995.

[79] Collins, F. S., Guyer, M. S., and Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**(5343):1580–1, 1997.

[80] Collins, F. S. Positional cloning: let's not call it reverse anymore. *Nature genetics*, **1**(1):3–6, 1992.

[81] Wooster, R., et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**(6559):789–92, 1995.

[82] Lifton, R. P. Molecular genetics of human blood pressure variation. *Science (New York, NY)*, **272**(5262):676–80, 1996.

[83] Yamagata, K., et al. Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature*, **384**(6608):455–8, 1996.

[84] Altshuler, D., Daly, M. J., and Lander, E. S. Genetic mapping in human disease. *Science (New York, NY)*, **322**(5903):881–8, 2008.

[85] Saiki, R. K., et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, NY)*, **230**(4732):1350–4, 1985.

[86] Saiki, R. K., et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, NY)*, **239**(4839):487–91, 1988.

[87] Yandell, D. W. and Dryja, T. P. Detection of DNA sequence polymorphisms by enzymatic amplification and direct genomic sequencing. *American journal of human genetics*, **45**(4):547–55, 1989.

[88] Antonarakis, S. E., Kazazian, H. H., and Orkin, S. H. DNA polymorphism and molecular pathology of the human globin gene clusters. *Human Genetics*, **69**(1):1–14, 1985.

[89] Orita, M., et al. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **86**(8):2766–70, 1989.

[90] Kruglyak, L. and Nickerson, D. A. Variation is the spice of life. *Nature genetics*, **27**(3):234–6, 2001.

[91] Reich, D. E., Gabriel, S. B., and Altshuler, D. Quality and completeness of SNP databases. *Nat Genet*, **33**(4):457–458, 2003.

[92] Fodor, S. P., et al. Light-directed, spatially addressable parallel chemical synthesis. *Science (New York, NY)*, **251**(4995):767–73, 1991.

[93] Chee, M., et al. Accessing Genetic Information with High-Density DNA Arrays. *Science*, **274**(5287):610–614, 1996.

[94] Lander, E. S. The new genomics: global views of biology. *Science (New York, NY)*, **274**(5287):536–9, 1996.

[95] Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. *Science (New York, NY)*, **273**(5281):1516–7, 1996.

[96] Yu, A., et al. Comparison of human genetic and sequence-based physical maps. *Nature*, **409**(6822):951–3, 2001.

[97] Broman, K. W., et al. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American journal of human genetics*, **63**(3):861–9, 1998.

[98] The International HapMap Consortium. The International HapMap Project. *Nature*, **426**(6968):789–96, 2003.

[99] Goldstein, D. B. and Cavalleri, G. L. A haplotype map of the human genome. *Nature*, **437**(7063):1299–320, 2005.

[100] Voight, B. F., et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*, **42**(7):579–89, 2010.

[101] Chen, W., et al. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration.

Proceedings of the National Academy of Sciences of the United States of America, **107**(16):7401–6, 2010.

[102] Franke, A., et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, **42**(12):1118–25, 2010.

[103] Lango Allen, H., et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**(7317):832–8, 2010.

[104] Maher, B. Personal genomes: The case of the missing heritability. *Nature*, **456**(7218):18–21, 2008.

[105] Goldstein, D. B. Common genetic variation and human traits. *The New England journal of medicine*, **360**(17):1696–8, 2009.

[106] Manolio, T. A., et al. Finding the missing heritability of complex diseases. *Nature*, **461**(7265):747–53, 2009.

[107] Gibson, G. Rare and common variants: twenty arguments. *Nature reviews Genetics*, **13**(2):135–45, 2011.

[108] Levy, S., et al. The diploid genome sequence of an individual human. *PLoS biology*, **5**(10):e254, 2007.

[109] Wang, J., et al. The diploid genome sequence of an Asian individual. *Nature*, **456**(7218):60–5, 2008.

[110] Wheeler, D. A., et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189):872–6, 2008.

[111] Pushkarev, D., Neff, N. F., and Quake, S. R. Single-molecule sequencing of an individual human genome. *Nature biotechnology*, **27**(9):847–50, 2009.

[112] 1000 Genomes Project Consortium, et al. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–73, 2010.

[113] Pearson, T. A. and Manolio, T. A. How to interpret a genome-wide association study. *JAMA : the journal of the American Medical Association*, **299**(11):1335–44, 2008.

[114] Duggal, P., et al. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC genomics*, **9**:516, 2008.

[115] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature*, **470**(7333):187–97, 2011.

[116] Kathiresan, S., et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, **40**(2):189–97, 2008.

[117] Burkhardt, R., et al. Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arteriosclerosis, thrombosis, and vascular biology*, **28**(11):2078–84, 2008.

[118] Istvan, E. S. and Deisenhofer, J. Structural mechanism for statin inhibition of HMG-CoA reductase. *Science (New York, NY)*, **292**(5519):1160–4, 2001.

[119] Tanner, M. M., et al. Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. *Cancer research*, **56**(15):3441–5, 1996.

[120] Pinkel, D., et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, **20**(2):207–11, 1998.

[121] Lucito, R., et al. Genetic analysis using genomic representations. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(8):4487–92, 1998.

[122] Pollack, J. R., et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature genetics*, **23**(1):41–6, 1999.

[123] Lucito, R., et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome research*, **10**(11):1726–36, 2000.

[124] Lisitsyn, N. and Wigler, M. Cloning the differences between two complex genomes. *Science (New York, NY)*, **259**(5097):946–51, 1993.

[125] Lucito, R., et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome research*, **13**(10):2291–305, 2003.

[126] Sharp, A. J., et al. Segmental duplications and copy-number variation in the human genome. *American journal of human genetics*, **77**(1):78–88, 2005.

[127] Redon, R., et al. Global variation in copy number in the human genome. *Nature*, **444**(7118):444–454, 2006.

[128] McCarroll, S. A., et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, **40**(10):1166–74, 2008.

[129] Shen, F., et al. Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC genetics*, **9**:27, 2008.

[130] Lim, S. A. *Single Molecule Systems: Advancements and Applications to Microbial and Human Genome Analysis*. Ph.D. thesis, University of Wisconsin-Madison, 2004.

[131] Reslewic, S. *The Optical Mapping of Genomes: Gaining New Insignts on Genome Structure and Variation by Single DNA Molecule Analysis*. Ph.D. thesis, The University of Wisconsin – Madison, 2005.

[132] Tuzun, E., et al. Fine-scale structural variation of the human genome. *Nature genetics*, **37**(7):727–32, 2005.

[133] Kidd, J. M., et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191):56–64, 2008.

[134] Korbel, J. O., et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, NY)*, **318**(5849):420–6, 2007.

[135] Fullwood, M. J., et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, **19**(4):521–32, 2009.

[136] Korbel, J. O., et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, **10**(2):R23, 2009.

[137] Hajirasouliha, I., et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics (Oxford, England)*, **26**(10):1277–83, 2010.

[138] Bailey, J. A. and Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics*, **7**(7):552–64, 2006.

[139] Le Maréchal, C., et al. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nature genetics*, **38**(12):1372–4, 2006.

[140] Hollox, E. J., et al. Psoriasis is associated with increased beta-defensin genomic copy number. *Nature genetics*, **40**(1):23–5, 2008.

[141] Pollex, R. L. and Hegele, R. A. Genomic copy number variation and its potential role in lipoprotein and metabolic phenotypes. *Current opinion in lipidology*, **18**(2):174–80, 2007.

[142] Fanciulli, M., et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, **39**(6):721–3, 2007.

[143] Ledbetter, D. H., et al. Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *The New England journal of medicine*, **304**(6):325–9, 1981.

[144] Girirajan, S., et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature genetics*, **42**(3):203–9, 2010.

[145] Amir, R. E., et al. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature genetics*, **23**(2):185–8, 1999.

[146] Neill, N. J., et al. Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. *Genome research*, **21**(4):535–44, 2011.

[147] Rovelet-Lecrux, A., et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature genetics*, **38**(1):24–6, 2006.

[148] Sebat, J., et al. Strong association of de novo copy number mutations with autism. *Science (New York, NY)*, **316**(5823):445–9, 2007.

[149] Stone, J. L., et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**(7210):237–41, 2008.

[150] Stefansson, H., et al. Common variants conferring risk of schizophrenia. *Nature*, **460**(7256):744–7, 2009.

[151] McCarthy, S. E., et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nature genetics*, **41**(11):1223–7, 2009.

[152] Singleton, A. B., et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science (New York, NY)*, **302**(5646):841, 2003.

[153] Helbig, I., et al. 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature genetics*, **41**(2):160–2, 2009.

[154] Erez, A., et al. Alu-specific microhomology-mediated deletions in CDKL5 in females with early-onset seizure disorder. *Neurogenetics*, **10**(4):363–9, 2009.

[155] van Ommen, G.-J. B. Frequency of new copy number variation in humans. *Nature genetics*, **37**(4):333–4, 2005.

[156] Perry, G. H., et al. Hotspots for copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(21):8006–11, 2006.

[157] Ciccarelli, F. D., et al. Complex genomic rearrangements lead to novel primate gene function. *Genome research*, **15**(3):343–51, 2005.

[158] Taylor, J. S. and Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annual review of genetics*, **38**:615–43, 2004.

[159] Stefansson, H., et al. A common inversion under selection in Europeans. *Nature genetics*, **37**(2):129–37, 2005.

[160] Schwartz, D. C. and Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, **37**(1):67–75, 1984.

[161] Morozov, V. N., et al. New polyacrylamide gel-based methods of sample preparation for optical microscopy: immobilization of DNA molecules for optical mapping. *Journal of microscopy*, **183**(Pt 3):205–14, 1996.

[162] Reed, J., et al. A quantitative study of optical mapping surfaces by atomic force microscopy and restriction endonuclease digestion assays. *Analytical biochemistry*, **259**(1):80–8, 1998.

[163] Qi, R. *Whole genome shotgun optical mapping of the Deinococcus radiodurans and Trypanosoma brucei genomes : and the development of a new surface system for optical mapping.* Ph.D. thesis, New York University, 1999.

[164] Jendrejack, R. M., et al. DNA dynamics in a microchannel. *Physical review letters*, **91**(3):38,102, 2003.

[165] Aston, C., Hiort, C., and Schwartz, D. C. Optical mapping: an approach for fine mapping. *Methods in enzymology*, **303**:55–73, 1999.

[166] Zhou, S., et al. A whole-genome shotgun optical map of Yersinia pestis strain KIM. *Applied and environmental microbiology*, **68**(12):6321–31, 2002.

[167] Zhou, S., et al. Whole-genome shotgun optical mapping of Rhodobacter sphaeroides strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome research*, **13**(9):2142–51, 2003.

[168] Anantharaman, T. S., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. II: Ordered restriction maps. *Journal of computational biology : a journal of computational molecular cell biology*, **4**(2):91–118, 1997.

[169] Anantharaman, T., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. III: Contiging genomic DNA. In *Proc Int Conf Intell Syst Mol Biol*, Proceedings 7th Intl. Cnf. on Intelligent Systems for Molecular Biology, pages 18–27. 1999.

[170] Sarkar, D. *Analyzing Optical Mapping Data using in silico Restriction Maps*. Ph.D. thesis, University of Wisconsin, 2005.

[171] Sarkar, D. *On the Analysis of Optical Mapping Data*. Ph.D. thesis, University of Wisconsin – Madison, 2006.

[172] Valouev, A., et al. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, **22**(10):1217–24, 2006.

[173] Valouev, A., et al. Alignment of optical maps. *Journal of computational biology : a journal of computational molecular cell biology*, **13**(2):442–62, 2006.

[174] Valouev, A. *Shotgun Optical Mapping: A Comprehensive Statistical and Computational Analysis*. Ph.D. thesis, University of Southern California, 2006.

[175] Valouev, A., et al. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(43):15,770–5, 2006.

[176] Li, H., et al. A quantile method for sizing optical maps. *Journal of computational biology : a journal of computational molecular cell biology*, **14**(3):255–66, 2007.

[177] Church, D. M., et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, **7**(5):e1000,112, 2009.

[178] Sarkar, D., et al. Statistical significance of optical map alignments. *Journal of computational biology : a journal of computational molecular cell biology*, **19**(5):478–92, 2012.

[179] Meng, X., et al. Optical mapping of lambda bacteriophage clones using restriction endonucleases. *Nature genetics*, **9**(4):432–8, 1995.

[180] Giacalone, J., et al. Optical mapping of BAC clones from the human Y chromosome DAZ locus. *Genome research*, **10**(9):1421–9, 2000.

[181] Jing, J., et al. Optical mapping of Plasmodium falciparum chromosome 2. *Genome research*, **9**(2):175–81, 1999.

[182] Lai, Z., et al. A shotgun optical map of the entire Plasmodium falciparum genome. *Nature genetics*, **23**(3):309–13, 1999.

[183] Perna, N. T., et al. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, **409**(6819):529–33, 2001.

[184] Zhou, S., et al. Shotgun optical mapping of the entire Leishmania major Friedlin genome. *Molecular and biochemical parasitology*, **138**(1):97–106, 2004.

[185] Schnable, P. S., et al. The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, NY)*, **326**(5956):1112–5, 2009.

[186] Wei, F., et al. The physical and genetic framework of the maize B73 genome. *PLoS genetics*, **5**(11):e1000,715, 2009.

[187] Wei, F., et al. Detailed analysis of a contiguous 22-Mb region of the maize genome. *PLoS genetics*, **5**(11):e1000,728, 2009.

[188] Ananiev, G. E., et al. Optical mapping discerns genome wide DNA methylation profiles. *BMC molecular biology*, **9**:68, 2008.

2 OPTIMIZATION OF OPTICAL MAPPING PROTOCOLS

The central goal of this thesis is to better understand the character and population distribution of normal germ-line variation in the structure of human genomes. Given how little is known about the occurrence and form of genome structure polymorphism [1–7], I undertook to survey structural variation in a set of phenotypically normal human genomes using Optical Mapping. Our interests in designing this study were twofold: first, we wanted to discover and characterize new structural variants, hypothesizing that Optical Mapping's scope, resolution and freedom from ascertainment bias should increase our power to do so over that of contemporary methods. Second, we were interested in studying structural variants in the context of other genomic landmarks [1], hypothesizing that these relationships could result in clues to their biological meaning.

Before performing any kind of analysis, though, I had to generate the Optical Mapping data sets to analyze. The success of Optical Mapping is predicated on the reproducible production, manipulation, presentation and analysis of ensembles of megabase-size molecules of DNA; these processes had been shown successful for prokaryotic and lower eukaryotic genomes [8–14], and a number of individual higher eukaryotes and mammalian and human samples had been analyzed [15–18]. However, at the time this project was conceived the Optical Mapping platform had not yet been applied to *populations* of mammalian genomes. Doing so in a reasonable time period required better data quality and higher throughput to leverage the emerging computational methods being developed at the time [17, 19–27].

Thus, I undertook a careful survey of the methods with which we produce, manipulate and present megabase-size molecules of DNA, with a particular eye towards

improving consistency and reproducibility. This chapter describes the experiments and analyses I performed in the course of optimizing pre-existing protocols or developing new ones, as well as detailing the final protocols themselves. Then, I present an account of the selection of several normal human samples and the generation of Optical Mapping data sets from them. Presentation of the subsequent data analyses is deferred until the next chapter.

2.1 Improved Methods for Optical Mapping

Introduction: An Overview of Optical Mapping Methods

The experimental protocols for generating Optical Mapping data are outlined in Figure 2.1. The protocols for generating Optical Mapping data have three prerequirements: first, long, pure genomic DNA is isolated from the sample under study. The isolation protocol is chosen or developed based on the organism being mapped, with the goal of keeping the DNA intact while freeing it from its encompassing proteins and lipid membranes. Second, Optical Mapping surfaces [28, 29] are prepared by treating microscope coverslip glass with a series of strong acid washes, cleaning the glass thoroughly and protonating the bulk silica matrix to create terminal hydroxyl groups on its surface. The coverslips are then incubated in an aqueous solution of organosilanes, which impart a positive surface charge to the cleaned glass. Finally, disposable microfluidic devices consisting of an array of microchannels [30] are fabricated by replica molding of poly(dimethylsiloxane) on a silicon wafer master patterned with photoresist features [31].

To generate single-molecule restriction maps (Figure 2.2), a microfluidic device is

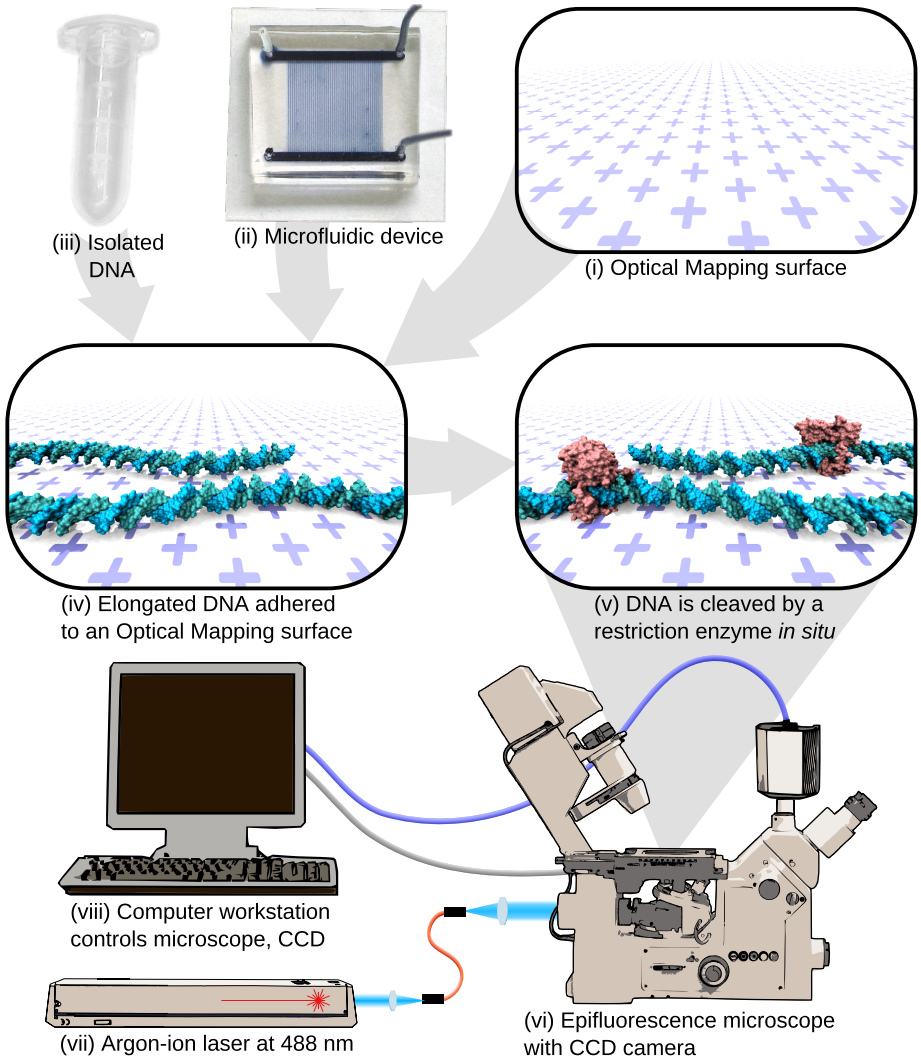

Figure 2.1. An overview of the Optical Mapping platform. Bulk microscope cover glass is cleaned with a Nano-Strip and 14 N hydrochloric acid, then treated with a silane mixture to make positively charged Optical Mapping surfaces (i). A silicon wafer is patterned with standard photolithography techniques, and then replicated into a flexible PDMS microfluidic device (ii) using soft lithography. Finally, pure, high molecular-weight DNA (iii) is isolated from cultured eukaryotic cells using a gentle detergent-based lysis protocol. The microfluidic device is adhered to the Optical Mapping surface, and the DNA solution is pumped through the microchannels, wherein the DNA is elongated and attached to the Optical Mapping surface *via* electrostatic interaction (iv). The DNA is incubated with a restriction endonuclease (v), which cleaves the DNA at its cognate sites. The cleaved DNA is stained and imaged on an epifluorescence microscope (vi) illuminated by an argon-ion laser (vii) and controlled by a computer workstation (viii).

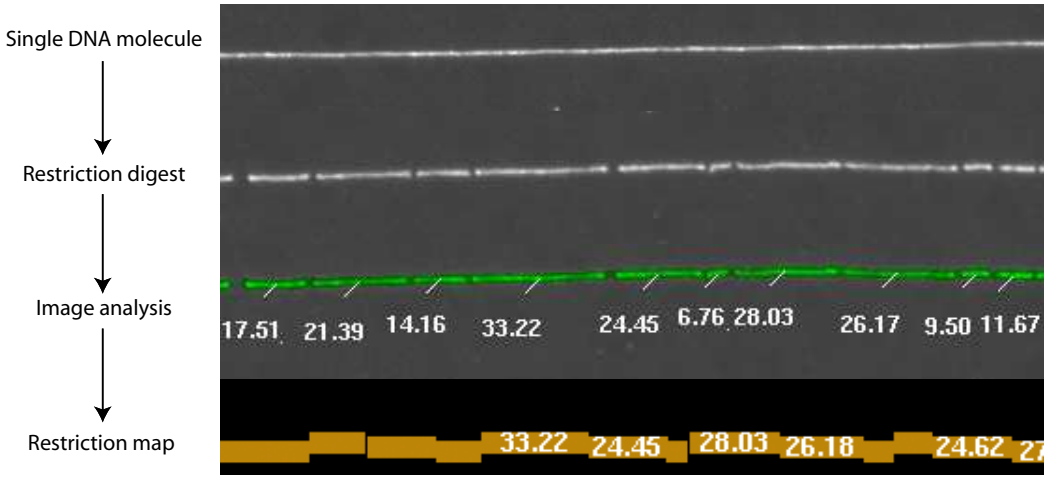

Figure 2.2. Basics of Optical Mapping. Optical Mapping generates ordered restriction maps from single molecules of genomic DNA.

adhered to a positively charged glass coverslip and the DNA solution flowed through the channels, depositing the negatively-charged DNA on the surface. Thus adhered and presented, the DNA molecules are further immobilized by the polymerization of a thin layer of polyacrylamide gel, which is sandwiched between the DNA-bearing Optical Mapping surface and a microscope slide and allowed to polymerize. After the surface is peeled off the slide, it is incubated with a restriction endonuclease which cleaves the DNA *in situ*, then rinsed and stained with the intercalating fluorescent dye YOYO-1 (Invitrogen, Carlsbad, CA). The stained surface is mounted on a microscope slide and place on an automated epifluorescence microscopy workstation, which images the stained DNA for analysis by custom-written machine vision software named PathFinder [30]. The machine vision software estimates each fragment's size based on its integrated fluorescence intensity, then groups together co-linear fragments and reports them as ordered restriction maps of single molecules of DNA.

Thus, there are many separate steps involved in producing Optical Mapping data from a biological sample. What follows is a detailed optimization of many of these

steps with an eye towards improving their reliability and reproducibility.

DNA Isolation from Mammalian Cells

The primary criterion for a DNA isolation protocol to be used with Optical Mapping is that it must free the DNA from the membranes and proteins in which it is packed with a minimum of breakage; that is, the resultant DNA must be long and pure. As a secondary criterion, it is nice that the resultant DNA solution be relatively homogeneous: a 3 µl aliquot taken from the top of the microcentrifuge tube should be about the same concentration as a 3 µl aliquot from the bottom of the tube. Since large DNA molecules have a negligible rate of diffusion [32, 33] this requires either post-purification trituration (and possible DNA breakage), or maintaining even dispersion of the cells being lysed.

When preparing mammalian cells for mapping, one common approach is to mix the cells with 1 mg/ml proteinase K in TE buffer (10 mM Tris buffer, 1 mM EDTA, pH 8.0), then incubate at 75 °C to lyse the cells [17]. Another approach uses a lysis solution composed of 1 mg/ml proteinase K, 10 mM EDTA, 1 mM EGTA, 20 mM NaCl and 0.01 mM spermidine, then calls for an hour's incubation at 65 °C followed by overnight incubation at 37 °C . Both have been used for successful Optical Mapping projects [17, 34], but their performance varies significantly from researcher to researcher.

When considering the refinement or *de novo* design of a DNA isolation protocol for Optical Mapping, it makes sense to consider the physical and chemical processes involved. The first requirement is to break open the cells; because mammalian cells don't have a cell wall, a little detergent should be sufficient to take care of the

lipid membranes. Other laboratory protocols, including those for the preparation of mammalian DNA in agarose gel inserts [35], call for N-lauroylsarcosine, an anionic detergent with a critical micellar concentration of 14.6 mM [36]. Usefully, it is available both as a free acid and as a sodium salt, allowing the experimenter to choose the form most suited for the pH of the solution being prepared. It also stays dissolved in aqueous solution at 4 °C instead of precipitating out.

The other critical process is the digestion of cellular proteins, particularly those that pack the DNA into chromatin. Protocols in the literature [35] and copious laboratory experience [9, 12, 13, 37–41] recommend proteinase K, a serine proteinase of broad specificity with several desirable properties. First, though it is activated by Ca^{+2} ions, they are not required for its catalytic activity and can be dispensed with for nucleic acid preparations [42]. Also, the enzyme is stable over a wide pH range, and is unaffected by a wide range of chelators and surfactants [43]. Finally, it can be inactivated by serine protease inhibitors [43] if its removal is required for downstream steps. It has maximum activity between 50 °C and 60 °C [43] and is frequently used at concentrations of 1 mg/ml [35].

Having settled upon reagents to effect the purification of mammalian DNA, what of its subsequent stabilization and protection from breakage? We consider three primary causes of DNA breakage. The first, and most problematic for long-term storage, is enzymatic cleavage by DNA nucleases. Nucleases are a common contaminant, and most rely on diavalent cationic cofactors to catalyze the cleavage reaction [44]. Thus, in addition to minimizing contamination by using proper sterile technique, long-term storage of DNA preparations frequently use moderate to high concentrations of EDTA to chelate and sequester the required cationic cofactors [35, 45, 46]. This also informs our selection of lysis pH, because EDTA's chelation efficiency depends on its

protonation state.

We also address the physical breakage of DNA by shear forces induced by manipulation, e.g. pipetting and vortexing. These concerns are addressed by reducing handling as much as possible: pipetting is minimized and only performed with a wide-mouthed pipette tip. Gentle trituration seem not to materially affect the quality of the preparation, but vortexing is to be avoided.

Finally, having considered some of the physical, chemical and enzymological requirements for a successful lysis, we turn to the assay by which we distinguish a good lysis from a bad one. A pulsed field gel electrophoresis (PFGE) [46] experiment is the traditional method for analyzing samples of megabase-size DNA; but running a PFGE gel takes days, and doing so at the sample concentrations used for Optical Mapping requires radioactive labeling.

In search of other methods that might speed up the protocol development cycle, I turned to the Optical Mapping literature and adapted from previous laboratory publications the following:

The sandwich mount [9, 12]. I placed an Optical Mapping surface on a glass microscope slide and pipetted a 5 µl drop of the DNA solution along one edge. Capillary action drew the solution into the space between the two, elongating the molecules. I then peeled them gently apart, allowing the surface to dry before staining with YOYO-1 and mounting on another microscope slide. This method resulted in many well-presented molecules, but they frequently elongated along the direction of the *peel* instead of in the direction of capillary flow, raising the possibility that they were torn in the process. For representative micrographs of these mounts, see Figure 2.3, top panels.

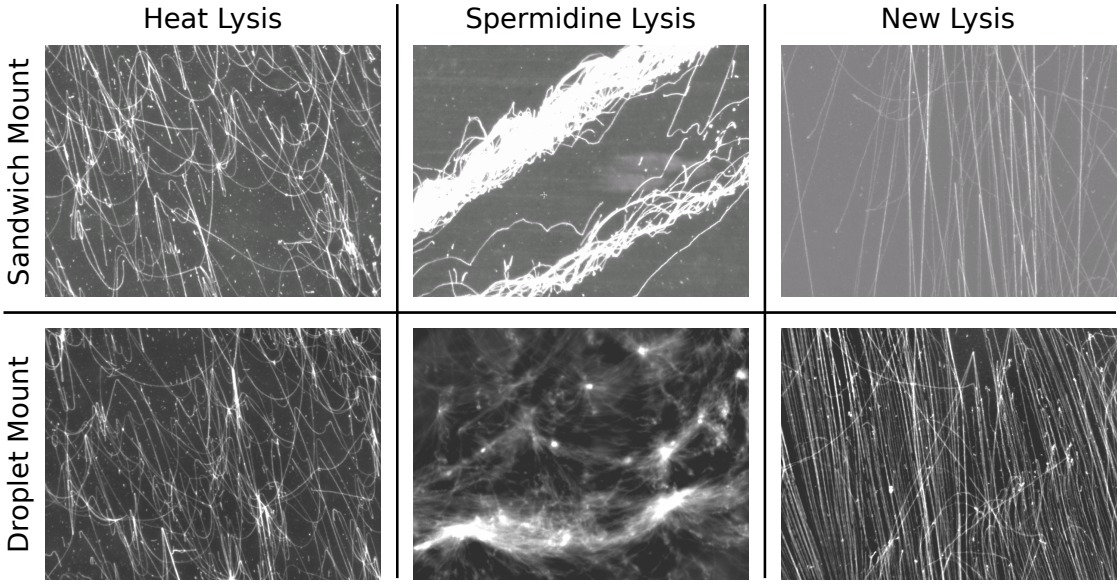

Figure 2.3. Lysis conditions effect DNA length, purity and homogeneity. These six panels compare three DNA preparation conditions using both the "sandwich mount" and "droplet mount" assays described in the text. The new lysis protocol results in complete lysis and results in a solution of long, pure DNA molecules: a good substrate for Optical Mapping.

The droplet mount [37]. I placed a 5 µl drop of the DNA solution on a dry Optical Mapping surface and allowed it to evaporate. The receding edge of the droplet elongated the DNA molecules, and it is clear that any DNA in the solution ends up on the surface; however, as the droplet dries, the solute conditions in the remaining liquid change. If these conditions affect DNA deposition, then that deposition is non-uniform. For representative micrographs of these mounts, see Figure 2.3, bottom panels.

Two months of investigation led to the lysis protocol described in Protocol 2.1. Representative micrographs of sandwich and droplet mounts are shown in Figure 2.3 alongside micrographs for the heat- and spermidine-based lysis protocols.

Protocol 2.1 Lysis protocol.

Solutions:

Phosphate-buffered saline (PBS)

- 137 mM NaCl
- 2.7 mM KCl
- 10 mM $Na_2HPO_4 \cdot 2H_2O$
- 2.0 mM KH_2PO_4
- pH 7.4

Lysis solution

- 20 mM EDTA pH 8.0
- 1 mg/ml proteinase K
- 10 mM N-lauroylsarcosine, sodium salt
- 2 mM Tris buffer pH 8.0

Procedure:

1. Pellet cells at 100xg, 4 °C in a swinging bucket centrifuge. Resuspend in the same volume of ice-cold PBS. Repeat once.
2. Count cells with a hemocytometer and dilute with ice-cold PBS to 2x the desired final concentration.
3. Mix 1:1 with ice-cold lysis solution. Incubate in a 50 °C water bath for 2 hours.
4. Cool to room temperature, then store at 4 °C .

Surface Chemistry: Preparation of Optical Mapping Surfaces

As with the lysis protocol optimization described above, I was guided in my investigation by a careful consideration of the physical and chemical processes involved in the attachment of silanes to hydrolyzed glass; these are sketched out in Figure 2.4. Most widely used organosilanes, including those used in the preparation of Optical Mapping surfaces, have one organic group and three hydrolyzable alkoxy groups bound to the central silicon atom. In the presence of water, the alkoxy groups hydrolyze to form

silanols, which then condense with (a) each other, to form silane polymers, or (b) with terminal hydroxy groups on the surface being treated to form a covalent linkage.

Standard laboratory practice [28, 29, 47] employs an aqueous mixture of N-trimethoxy-silylpropyl-N,N,N-trimethylammonium chloride and vinyltrimethoxysilane; the former imparts a positive surface charge, while the latter provides attachment sites for the polyacrylamide gel overlay, described later. In aqueous solution, the hydrolysis of the methoxy groups likely occurs as well as some amount of oligomerization. Acid-cleaned cover glass is incubated in this solution overnight, ostensibly to allow the covalent attachment of the silanol polymers to the glass substrate. The derivatized surfaces are stored in dry ethanol until immediately before use, then dried under ambient conditions before being assembled with a microfluidic device to present molecules.

A major stumbling block for the optimization of the protocol presented above was the lack of an easy assay for quantifying a surface's derivatization, particularly its macro-scale distribution over the surface [29, 48, 49]. If the silanization of the surface is uneven, it could complicate efforts to optimize downstream parameters affecting molecule deposition, digestion and staining: it is useful to treat all the DNA molecules on an Optical Mapping surface as having been subject to the same conditions, which is not the case if there is substantial spatial variation present in the derivatization. Unfortunately, direct measurement of silane deposition requires sophisticated analytical techniques beyond the scope of this project, and though the literature reveals several possible techniques, each has its drawbacks: X-ray photoelectron spectroscopy provides chemical information [50] but has limited spatial resolution. Spectrophotometric techniques require a chromophore, which rules out both of our commonly used silanes.

Fortunately, a number of indirect measurements are available to us, courtesy of the machine vision software (PathFinder [26, 30]) we use to measure molecules in micrographs. For each single molecule map it finds, PathFinder also reports a number of metrics related to the "quality" of the measurement. These metrics include average size of the gap between the molecule's fragments, the molecule's average stretch, and how many internal standards were used for estimating the fragments' sizes, among others.

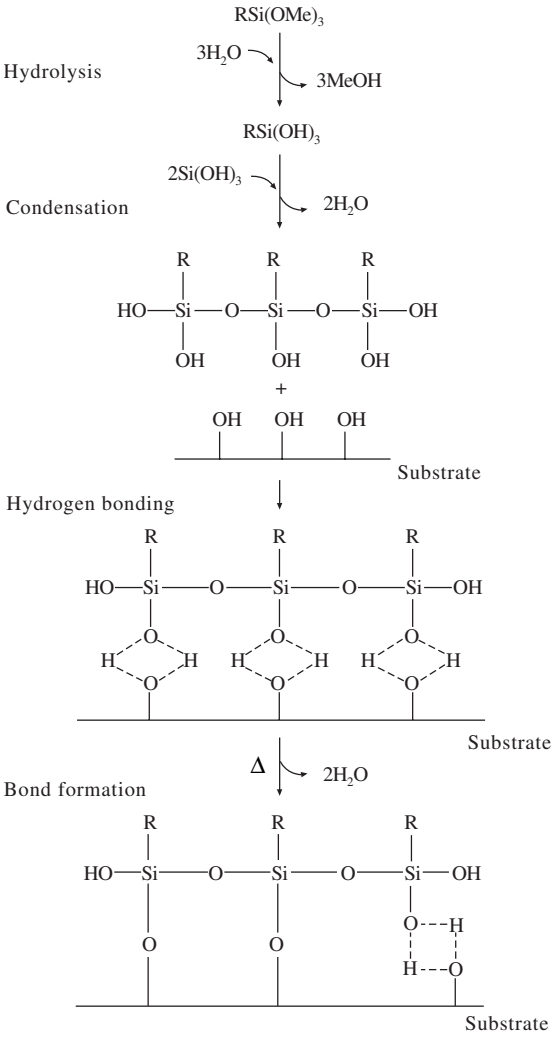

Figure 2.4. Hydrolytic deposition of silanes. Adapted from B. Arkles, CHEMTECH 7 p. 776 (1977).

How are these metrics useful? They're useful because they serve as proxies for physical attributes in which we're interested but are otherwise difficult to measure. For example, the visible gaps at digestion sites form because the DNA molecules are under tension and retract when they're cut by by the restrction endonuclease. The extent of this retraction depends directly on the magnitude and distribution of the forces binding the DNA molecule to the surface. Thus, the size of the gaps between fragments in a set of maps can be used as a proxy for measuring how "sticky" a surface is, a function of the positive charge imparted by the silanes that give it a positive charge. These

functional assays serve as imperfect, but useful, surrogates to impractical analytical methods.

A second useful feature of these quality metrics is that the location of the measurement on the surface is also saved. This allowed me to analyze not only the *magnitude* of a surface's "stickiness" (degree of derivatization) but also its *spatial distribution*. I employed a method called "kriging" [51], first developed in field of geostatistics [52], which estimates the spatial distribution of a continuous stochastic process given a set of observations from arbitrary locations in that field. It can be thought of as a spatial generalization of traditional linear least-squares regression, estimating the value of some unknown function f at a an arbitrary point $x^\star$ given the values of f at some other points $x_1 ... x_n$.

In this case, I have the average gap size of all the molecules PathFinder found on the surface and those molecules' locations; what I want to estimate is the spatial distribution of the stochastic process (i.e. surface derivatization) that led to those particular values. Examples of this analysis applied to several surfaces (as computed with the R package **RandomFields**) are presented in Figure 2.5. Each plot represents the spatial distribution of that surface's "stickiness" of a particular surface, as measured by molecule gap size.

As is apparent from Figure 2.5, different methods of derivatizing, storing and drying Optical Mapping surfaces have a significant impact on the properties of the surface for subsequent mapping. A month of exploring variations on the standard laboratory procedure resulted in the protocol presented in Protocol 2.2. (The acid cleaning steps remain unchanged, but are included for completeness' sake.)

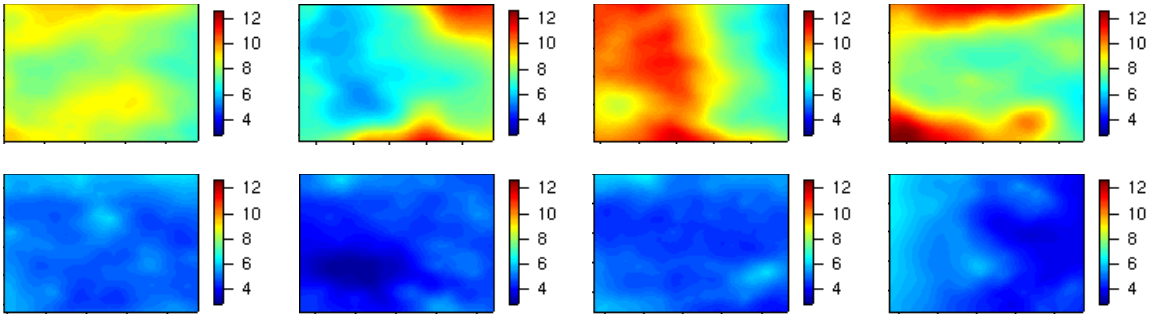

Figure 2.5. Drying conditions affect Optical Mapping surface properties. These eight subfigures each represent the restriction gap size distribution across one surface worth of collection, which serves as a proxy for that surface's stickiness. The top four were dried for 10 minutes under ambient conditions, and the bottom four were dried for 10 minutes at 65 °C on a digital hotplate. All eight surfaces were mounted under pump-driven fluid flow with identical flow parameters. Note that the hotplate-dried surfaces are not only sticker (smaller gap size) but more evenly so than those dried under ambient atmosphere.

Molecule Deposition: The Device

Despite many efforts to the contrary, the mechanisms behind molecule deposition and presentation remain some of the most poorly controlled in the entire Optical Mapping pipeline. In current laboratory practice [30], a microfluidic device composed of an array of microchannels is fabricated by first creating a negative master on a silicon wafer using standard photolithography techniques, then making replica molds of the pattern in poly(dimethylsiloxane) (PDMS) [30]. To mitigate PDMS' inherent hydrophobicity, a glassy layer is formed on the microchannels' surface by treatment with an oxygen plasma, and prevented from readsorbing into the bulk PDMS by storing the devices in distilled water until immediately before use. The devices are adhered to the Optical Mapping surface, and a drop of DNA solution is smeared along one side of the device, wetting the ends of the microchannels. The microchannels fill via capillary action, elongating and depositing the DNA molecules on the surface.

Protocol 2.2 Optical Mapping surface derivatization protocol.

Procedure:

- Load several ounces of FISHERfinest 22x22 mm cover glass (Fisher Scientific, Pittsburgh, PA) into the custom-built PTFE racks and seal in a reaction vessel. Submerge in Nano-Strip (Cyantek, Fremont, CA) and heat to 70 °C for one hour. Rinse glass under continuously running DI water until the effluent pH is greater than 7.0, as measured with a pH indicator strip.
- Submerge the cover glass in 14N hydrochloric acid (Fisher Scientific) and heat to 104 °C for six hours. Rinse glass under continuously running DI water until effluent pH is greater than 7.0, as measured with a pH indicator strip. Store submerged in absolute ethanol.
- Load 50 surfaces in an upright custom-machined PTFE rack in an HDPE jar (Qorpak, Bridgeville, PA).
- Mix 250 ml distilled water with 40 µl N-trimethoxy-silylpropyl-N,N,N-trimethyl-ammonium chloride (50% solution in methanol) and 10 µl vinyltrimethoxysilane (neat) (both from Gelest, Morrisville, PA) and shake vigorously for 60 seconds. Pour gently over the surfaces and incubate at 65 °C overnight with agitation.
- Rinse surfaces in the jar twice with distilled water, twice with absolute ethanol, and store under 85% ethanol until use.
- Immediately before use, remove the optical mapping surfaces from their storage solution and dry on a digital hot plate at to 65 °C for 10 minutes.

The remaining DNA sample is aspirated off, and the device is peeled off the surface, leaving the mounted DNA behind it.

The opportunities for variation and error in the above description are manifold. As per [30], the average velocity of the DNA solution in a microchannel is given by the Hagen–Poiseuille equation:

$$\bar{v} = \frac{C_g \Delta p}{\eta L}$$

where $\bar{v}$ is the average velocity, C_g is a geometric form factor derived from the dimensions of the channel, Δp is the pressure drop, η is the viscosity of the fluid and

L is the length of the channel. The pressure drop Δp can be described by capillary pressure P_c, which is calculated by

$$P_c = \gamma \left(\frac{cos(\theta_{substrate}) + cos(\theta_{PDMS})}{a} + \frac{2cos(\theta_{PDMS})}{b} \right)$$

where γ is the surface tension of the liquid, and $\theta_{substrate}$ and θ_{PDMS} are the contact angles of the Optical Mapping surface and PDMS channels, respectively.

While the physical dimensions of the channels C_g are not likely to change day-to-day, the viscosity of different DNA preparations can differ substantially depending on the protocol used to produce the solution, as can the properties of the Optical Mapping surface. These simplifications also don't take into account vagaries of the application and aspiration of the DNA solution. Thus, work was undertaken by myself and other laboratory members to mitigate these concerns and make DNA mounting more predictable, controllable and reproducible.

The result was a novel microfluidic device with the specifications outlined in Figure 2.6. It retains the array of microchannels, but places at their ends two reservoirs ("ports") that serve as interfaces between the micro-sized regime of the channels and the macro-manipulable world. In usual practice, holes were cored through the bulk PDMS into both sides of each port, allowing placement of Teflon tubing through which solutions could be pumped into the ports and thence into the microchannels. Attached to syringe pumps controlled and coordinated by computer software I wrote, this setup allows precise and reproducible control of fluid flow through the microchannels. Its fabrication is detailed in Protocol 2.3.

The goal of such precise, reproducible control over molecule deposition and presentation was to enable the creation of a feedback loop, with the results ("quality")

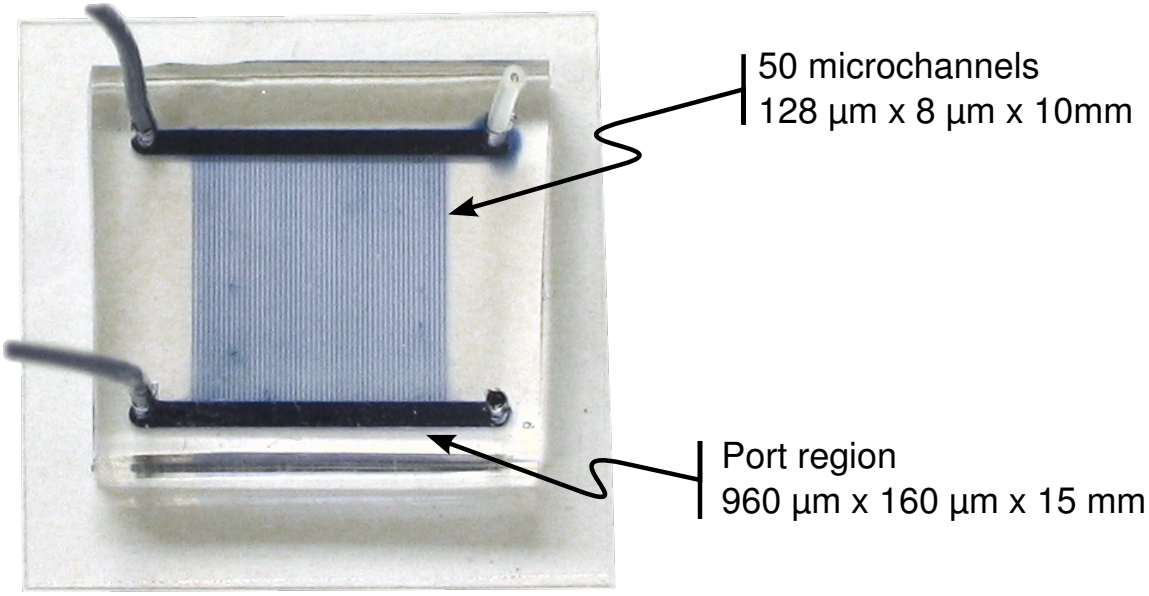

Figure 2.6. A microfluidic device for pump-mediated molecule deposition. Included are the relevant feature dimensions.

Protocol 2.3 Microfluidic device fabrication

---

1. Five hundred micron silicon wafers (Montco Silicon Technologies, Spring City, PA) were patterned using standard photolithographic techniques and an SU-8 negative photoresist (MicroChem Corp., Newton, MA).

2. The pattern is an array of microcapillary channels, each of dimensions 128 µm by 8 µm by 10 mm. At each end of the array of microchannels is a larger "port" region, each with dimensions 960 µm by 160 µm by 15 mm (Figure 2.6).

3. Mix Sylgard 184 silicone elastomeric base (Dow Corning, Midland, MI) with its curing reagent as per the Sylgard product sheet. Pour over the patterned wafer and cure at 65 °C overnight.

4. Peel the sheet of elastomer off of the patterned silicon wafer and cut apart the microfluidic devices with a clean razor blade.

5. Core four holes (one at each end of both ports) using a 20-gauge needle, to provide access to the microfluidic channels. Needles must be clean of lubricant (which is highly fluorescent!); methanol and a lens paper are effective at removing it.

---

of one experiment informing the choice of parameters of the next. Besides the ability to reproducibly control molecule deposition, two additional tools were needed. First, as with the surface dervitization optimization described above, we need metrics to track how changes in loading protocols affect the molecules' presentation. Again, we turn to the quality metrics produced by our machine vision software, PathFinder [26, 30], for this purpose.

The other necessary component is an *extrinsic* measure of molecule quality: how *informative* is the map derived from a molecule? How useful is it? Such an extrinsic measure provides the "objective function" for our protocol optimization: we want to change the protocol in a stepwise fashion such that whatever measurement we choose is maximized.

And, it turns out, we already have one ready-at-hand: SOMA pairwise alignments. SOMA ("Software for Optical Mapping Analysis") [26] is a toolkit of algorithms and utilities for the manipulation and analysis of Optical Mapping data sets. One of the most frequently used pieces of the SOMA toolbox is an optimal pairwise alignment tool based on the classic dynamic programming algorithms of Needleman and Wunsch [53] for global alignments and Smith and Waterman [54, 55] for local alignments. If we're mapping a genome (say, mouse or human) for which a high-quality reference sequence exists, and we further assume that most of the differences between our experimental genome and the reference sequence are small and localized, then whether or not a restriction map aligns to that reference is a crude measurement of that map's extrinsic quality. This assumption is so common that automatic SOMA alignments are a usual part of the Optical Mapping data collection pipeline.

The ready availability of both SOMA alignments and the PathFinder molecule metrics suggest the following strategy for closing the feedback loop. First, mount

molecules with an arbitrarily selected set of flow parameters. (In practice, the starting point was chosen to mimic as much as possible the flow rate in the capillary channels.) Digest and collect the data set as usual; then, divide the resultant molecules into those that aligned to the reference and those that did not. For each quality metric, construct a quantile-quantile (Q-Q) plot comparing the distribution of that metric in the molecules that aligned versus the distribution for those that did not. An example of such a "collection dashboard" tool's output is given in Figure 2.8. Based on the quality metric distributions, adjust flow parameters for the next collection: for example, if molecules with a higher estimated stretch were more likely to align to the reference, the next deposition might be performed with a higher flow rate through the microchannels. This procedure is iterated as necessary to maximize the proportion of molecules whose maps align to the reference. The pump flow regime on which I settled is included in Protocol 2.4.

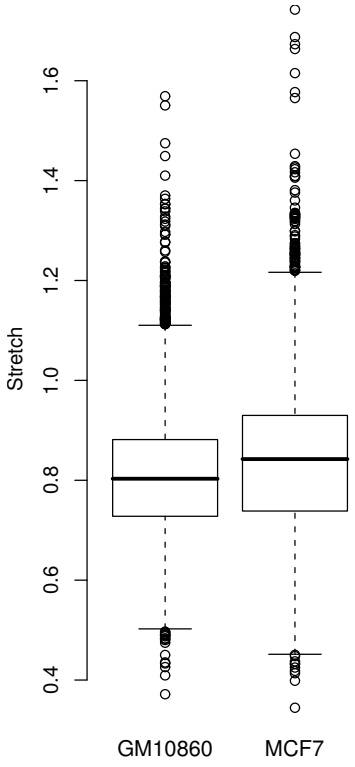

Figure 2.7. Molecule stretch compared between pump-mounted and capillary-mounted DNA molecules. Molecule stretch is less variable with pump-mounted DNA than with capillary-mounted DNA.

Finally, one would expect that if the pump-mediated loading described above really were more reproducible than capillary mount loading, there would be a decrease in variance in the distribution of some of the quality metrics described above. As Figure 2.7 demonstrates, that is indeed what we see when comparing the estimated stretch of DNA molecules from a pump-mounted collection (cell line GM10860, discussed in the next

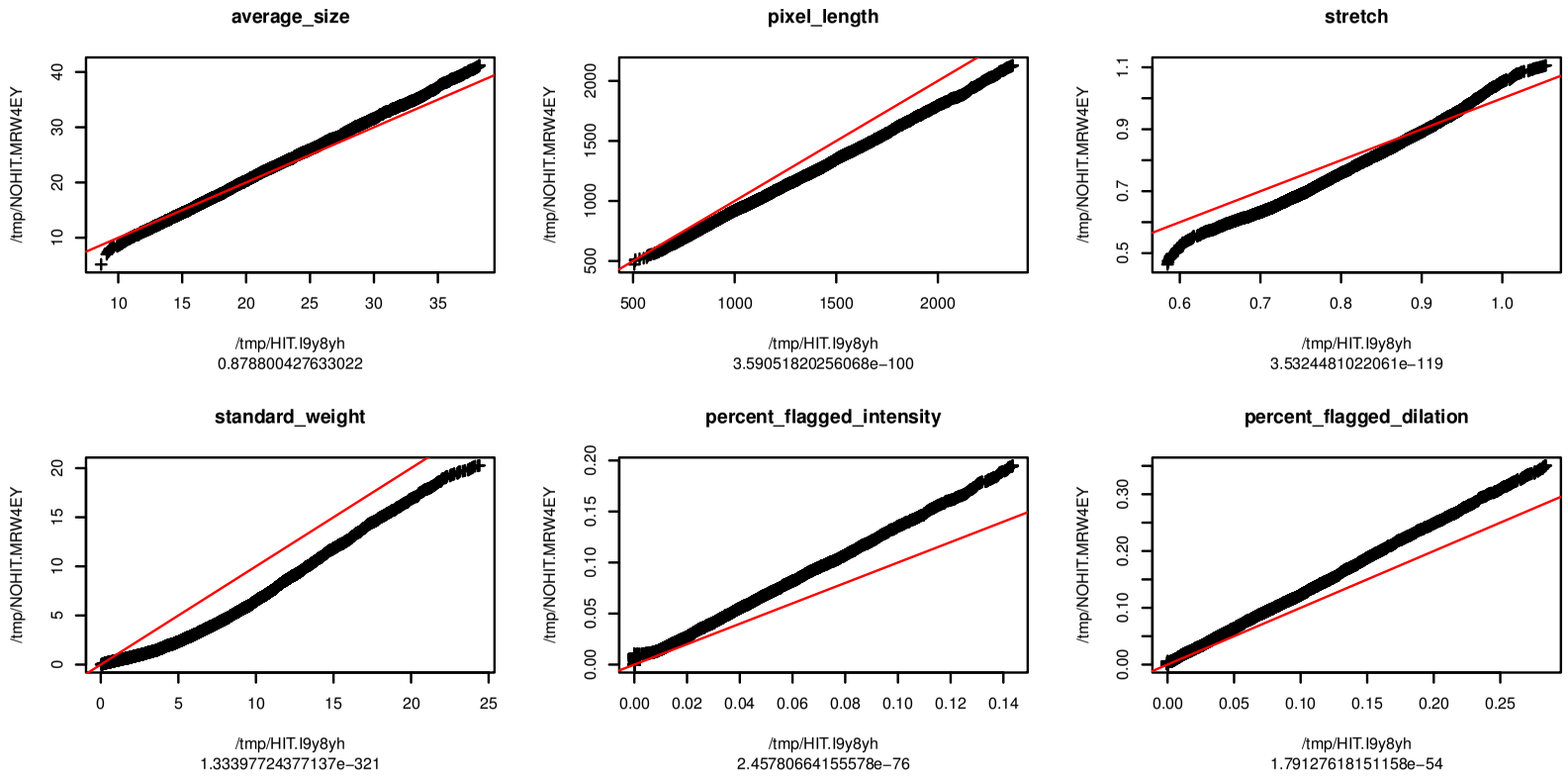

Figure 2.8. An example output of the collection dashboard. Each quantile-quantile plot compares the distribution of one of the quality metrics (above each plot) between maps that aligned back to the reference (X axis) and maps that didn't (Y axis.) Note in particular the substantial difference in the **standard_weight** metric, indicating an insufficiency of internal standards with which to estimate sizing. There are also the differences in the **percent_flagged_intensity** and **percent_flagged_dilation** metrics, both of which relate to evenness of staining. Staining is addressed in the next subsection.

Protocol 2.4 Molecule deposition protocol

- Adhere the microfluidic device to an Optical Mapping surface. Using a micropipette or syringe, fill the device with TE buffer.
- Connect lengths of teflon tubing to a 250 µl gastight syringe and a 5 µl microsyringe. Fill the both assemblies with TE buffer by removing the plunger and back-filling with a syringe or micropipette. Mount the large syringe on syringe pump 1, and the small syringe on pump 2.
- Connect the teflon tubing from syringe 2 to the top-left port hole in the device. Plug the top-right port hole with a piece of teflon tubing containing polymerized PDMS, to stopper up the hole.
- Using the syringe pump 1 (the large syringe), withdraw 0.7 µl air, then 6 µl of TE buffer, then 6 µl of DNA solution.
- Connect the end of the tubing attached to syringe 1 (containing the DNA sample backed by TE buffer) to the bottom-left port hole of the microfluidic device. Leave the bottom-right hole empty.
- Using the syringe pump control software, perform the following sequence of steps:

    1. Infuse the 6 µl of sample from syringe 1 into the bottom port at a rate of 5 µl/min . Simultaneously, infuse 0.12 µl of TE into the top port at a rate of 0.1 µl/min .
    2. Wait 60 seconds for DNA to resume a random-coil conformation.
    3. Using syringe 2, withdraw 0.3 µl at a rate of 0.5 µl/min to load sample into the channels.
    4. Change the syringe 2 withdrawal rate to 0.03 µl/min to deposit the sample. Withdraw an additional 0.1 µl .
    5. Infuse the 6 µl of TE from syringe 1 into the bottom port at a rate of 5 µl/min . Simultaneously, withdraw 0.6 µl from syringe 2 at a rate of 0.5 µl/min . This flushes the channels of remaining DNA.

- Proceed directly to the overlay protocol, below.

chapter) as compared with a project collected with the traditional capillary mount (MCF-7) [18].

Overlay Polymerization

I addressed overlay polymerization only briefly. As described in this section's Introduction, a polyacrylamide solution is polymerized on the Optical Mapping surface where it bonds covalently to the vinyl moieties of the silane polymer; the polyacrylamide pins down the DNA, helping to prevent small fragments from desorbing [13, 14, 30]. It also provides some distance between the surface and the slide on which it is mounted for microscopy, reducing background.

On the other hand, an overlay with overly small pores in the gel matrix might restrict access of the restriction enzyme to the DNA, slowing down or preventing digestion. Free acrylamide monomers might also interfere with enzymology or staining. In order to ensure controlled and reproducible characteristics of the gel overlay, it behooves the investigator to make sure that the polyacrylamide is fully polymerized before proceeding with mapping. BioRad specifies [56] that a traditional polyacrylamide preparation is fully polymerized after 60 minutes; how-

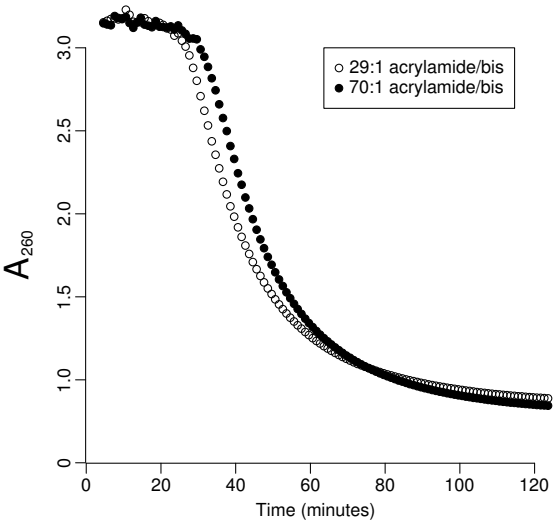

Figure 2.9. Acrylamide polymerization over time. Absorbance at 260 nm detects the acrylamide monomer's alkene double-bond, which is converted to an alkane bond by the polymerization reaction. Measured in a Hewlett Packard 8453 UV/Visible spectrophotometer, blanked with distilled water.

ever, because we use a very low concentration of acrylamide for the gel overlay, it is possible that the dynamics of that reaction are different. To measure them, I performed a polymerization reaction with standard Optical Mapping conditions (see Protocol 2.5) in the cuvette of a spectrophotometer, monitoring absorbance at 260 nm which measures the disappearance of the acrylamide monomer's double bond. As Figure 2.9 clearly shows, the polymerization reaction is essentially complete at 60 minutes. Protocol 2.5 records the entire overlay protocol.

Protocol 2.5 Optical Mapping overlay protocol.

Solutions:

Ammonium persulfate solution

- 10% w/v ammonium persulfate, made fresh in distilled H_2O.

Acrylamide solution

- 912.5 µl H_2O
- 82.5 µl 40% 29:1 acrylamide:bis solution
- 5 µl 2% Triton X-100, Molecular biology grade, Sigma
- Vortex to mix, then degas for 15 minutes under vacuum

Procedure:

1. Mix acrylamide solution, 7.5 µl ammonium persulfate solution and 0.8 µl TEMED. Vortex briefly.
2. Pipette 12 µl of the mixture onto the edge of the surface on which DNA has been mounted. Carefully lower surface onto a glass microscope slide, taking care to avoid bubbles.
3. Polymerize for 60 minutes in a humidity chamber (pipette tip box filled part-way with water).
4. Peel surface from slide with razor blade and place face-up in another humidity chamber for digestion. Proceed directly to the endonuclease restriction digest (Protocol 2.6).

DNA Restriction Digest

The precise restriction digest protocol depends on the enzyme being used, which in turn depends on the genome being digested. Protocol 2.6 was developed for digesting human DNA with the restriction endonuclease SwaI, the enzyme used to collection the normal human data sets described below.

DNA Staining

Because integrated fluorescence intensity is used to estimate the nucleotide content of each DNA fragment, consistent and reproducible staining is critical for the production of high-quality Optical Mapping data. I created a tool to compute metrics related to stain consistency from the PathFinder output files, including a molecule's mean signal-to-noise ratio (SNR); the variance of the SNR of the pixels along a molecule's backbone; local pixel intensity variation (based on a sliding window); and the proportion of the molecule that is "dim," i.e. has an intensity less than 70% of the molecule's mean intensity.

Using a Q-Q plot strategy similar to that described for the optimization of the pump loading, I examined the effect of these staining variables on map quality. Given that a fragment's size is computed directly from its fluorescence intensity, it comes as no surprise that the distributions of these metrics are different in maps that successfully aligned to the reference genome (high-quality maps) versus those maps that did not align (lower-quality maps.)(Figure 2.10).

As with the pump loading parameter optimization, these metrics and analyses formed the basis for a feedback loop with which I could optimize the staining protocol. The relative amounts of YOYO-1, the cyanine dimer stain we use, and β-

Protocol 2.6 DNA restriction digest protocol.

Solutions:

Bis-tris phosphate (BTP) buffer

- 10 mM BTP
- 100 mM NaCl
- 10 mM $MgCl_2$
- Adjust pH to 7.9. Filter-sterilize.

DTT solution

- 100 mM dithiothreitol
- Filter sterilize, store 1 ml aliquots frozen at -20 °C

Procedure:

1. Immediately after peeling the surface off the glass slide with its attached overlay (Protocol 2.5), rinse the surface with 500 µl filter-sterilized TE buffer, allowing each wash to sit on the surface for two minutes. After each wash, aspirate off the TE buffer.

2. Rinse with 200 µl BTP buffer without enzyme, waiting 2 minutes.

3. While the surface is being rinsed with BTP buffer, mix 200 µl BTP buffer with 2 µl 2% Triton X-100, 2 µl 100 mM DTT and 2 µl SwaI enzyme. Vortex briefly to mix. After 2 minutes, aspirate off the buffer rinse and apply the 200 µl digestion mixture.

4. Incubate at 25 °C for approximately 90 minutes, adjusting as necessary to achieve an 80% digestion rate as measured by average fragment size. For SwaI on a sample of human DNA, this means an average fragment size of about 21 kb.

5. Stop the reactions. Aspirate off the digestion buffer, and rinse three times with 500 µl filter-sterilized TE buffer as above.

6. Proceed directly to staining (Protocol 2.7).

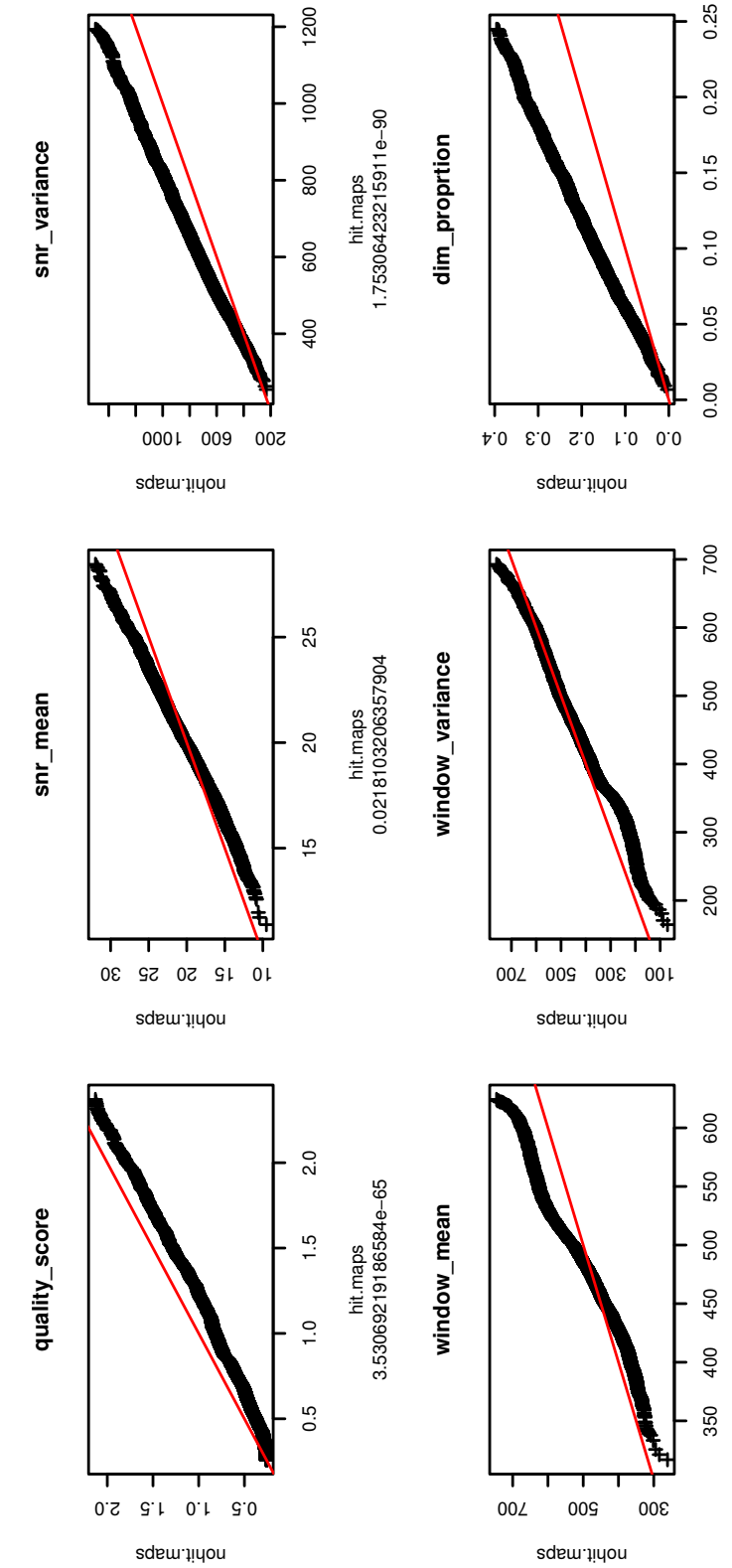

Figure 2.10. Dashboard view of staining parameters. Metrics include per-molecule SNR mean and variance, as well as the mean and variance of a local measure of staining variability and the proportion of the molecule that is dim, i.e. intensity less than 70% of the molecule mean intensity. There is relationship between several of these variables and data quality.

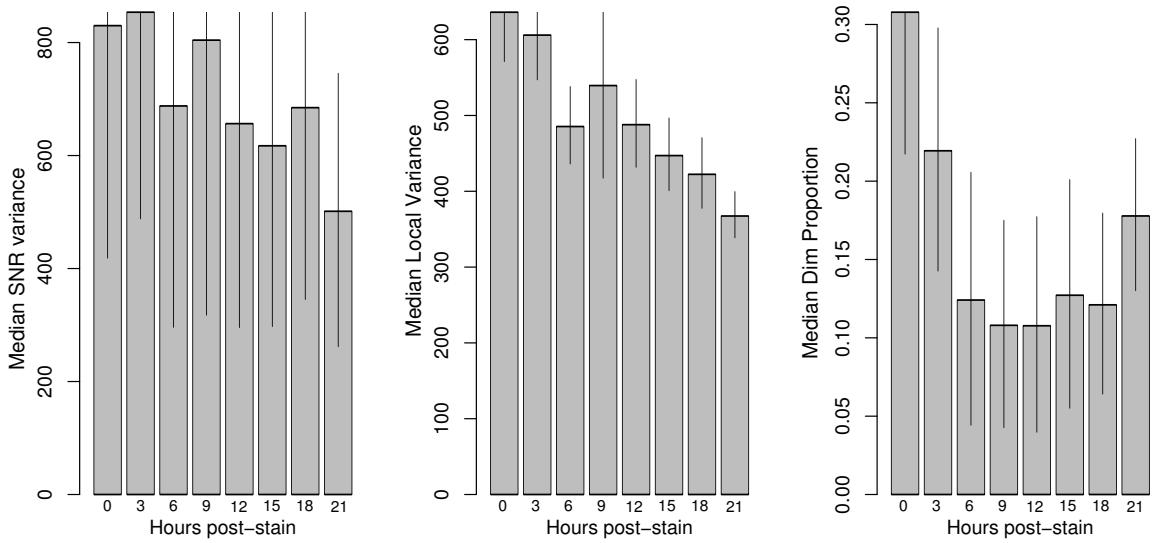

Figure 2.11. Variation in staining decreases with time. Both inter-molecule and intra-molecule measures of variation improve.

mercaptoethanol, the free-radical scavenger we use to slow photobleaching [57], don't appear to have a huge effect on data quality. What does seem to matter is staining the DNA, then allowing them to *sit overnight* before collection. As Figure 2.11 shows, the variance in SNR, mean local intensity variance and dim proportion all decrease as time progresses; and as per Figure 2.10, all three metrics have a direct impact on data quality. The complete staining protocol is given in Protocol 2.7.

Conclusion

The success of an Optical Mapping data collection project depends on the synergistic performance of many different protocols. As traditionally performed, a number of these protocols show worrisome dependence on ambient conditions, tight timing and the researcher's steady hands. The preceding sections record a concerted effort to optimize these protocols with an eye towards consistency and reproducibility; it

Protocol 2.7 YOYO-1 solutions and DNA staining protocol.

Solutions:

YOYO-1 100x stock solution

- To the 200 µl stock solution of YOYO-1 from Invitrogen, add 19.8 ml DMSO in a brown glass bottle.
- Filter using a Whatman 0.2 µm PVDF syringe filter.
- Pipette 50 µl aliquots into brown autoclaved 500 µl microcentrifuge tubes. Store at -80 °C .

YOYO-1 working solution

- Thaw a 50 µl aliquot of 100x stock YOYO-1 stock solution by vortexing at room temperature.
- Mix 4 ml TE pH 8 buffer, 1 ml B-mercaptoethanol, and the 50 µl YOYO-1 stock aliquot. Vortex well.
- Filter-sterilize with a 0.2 µm polypropylene syringe filter into black light-safe microcentrifuge tubes. Store at 4 °C .

Procedure:

1. After aspirating the last post-digestion TE rinse from the Optical Mapping surface, allow it to dry in its humidity chamber for 10 minutes.
2. Apply 12 µl of YOYO-1 working solution to the optical mapping surface. Carefully invert and lower it slowly onto a microscope slide, allowing air bubbles to escape.
3. Seal the wet-mounted surface to the microscope slide with two coats of nail varnish.
4. Wait at least 12 hours before imaging.

is hoped that, in addition to the optimized protocols themselves, the procedures and analytical tools described in the narrative will be useful in further protocol development efforts.

2.2 Optical Mapping of Three Normal Human Genomes

Introduction

Their goal of the protocol optimizations described in the previous section was to enable the Optical Mapping platform to analyze larger numbers of samples, in order to begin to address populations instead of being constrained to one-off samples. This section describes using the new protocols to collect Optical Mapping data sets from three normal human genomes: the lymphoblast-derived cell lines GM15510, GM10860 and GM18994. Their selection, culture and mapping is described below.

Sample Selection

We chose samples to analyze based on two criteria. First is the availability of additional genomic data against which to compare our findings. For this criteria, no sample compares to the cell line GM15510 from the National Institute of General Medical Sciences' Polymorphism Discovery Resource. The GM15510 cell line was the source of the G248 fosmid clone library, which was created and end-sequenced as part of the effort to finish the human genome reference sequence [58]. These end-sequencing data were subsequently used by Tuzun et al. [59] in one of the first analyses of fine-scale human structural variation; since then, it's been the subject of microarray-based

copy number analysis [60] as well as several sequencing-based studies [61–63]. The wealth of preexisting data provides for a direct comparison of our results with theirs, allowing us to evaluate relative strengths and weaknesses and situate our results in the context of other current approaches.

GM15510 (seed)
GM10860
GM18994
GM18505
GM11839
GM18501
GM19202
GM19203
GM18515
GM19145
GM19206
GM18857
GM19210
GM19161
GM18912
GM19141

Table 2.1. The top 15 most diverse HapMap samples, as determined by SNP genotypes.

Our second criterion is maximum diversity: if we're limited to analyzing only a handful of samples, it would be nice to maximize the number of variants we discover. How can we choose samples that are maximally different as to provide the most diversity of polymorphism to study? Race makes a poor proxy for biological heterogeneity [64, 65], but the wide availability of SNP genotype data provides a rational basis on which to make our selection, assuming that diversity in SNP genotype serves as a proxy for genetic diversity overall [66]. Thus, the goal becomes choosing samples whose SNP genotypes are maximally different for some definition of "different." I chose the Hamming distance [67] to measure the pairwise difference between two samples, because it is easy to compute and satisfies the axioms for a metric. For SNP genotypes, it is simply the number of loci at which the two samples have a different genotype. The sample set from which I chose was the 269 samples in the HapMap panel [68–70]: not only does each sample have over a million high-quality SNP genotypes, but the actual DNA samples and cell lines are commercially available for additional analysis. How exactly to go about choosing a maximally diverse set of HapMap samples? Given a set of cell lines D and

Algorithm 2.1 The $MaxMin$ algorithm.

Require: $D \neq 0$
Require: $n \neq 0$
Require: $s \in D$
  $S \leftarrow s$
  while $D < n$ **do**
    Choose d from D maximally distant from S.
    $S \leftarrow d$
  end while
  return S

a metric d_{ij} that measures the difference between samples i and j (their Hamming distance, above), I propose that a subset $S \subset D$ is maximally diverse if if maximizes the sum of the distance between the members of S. In particular, it maximizes

$$Z = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} d_{ij}$$

While the maximum-diversity problem as formulated above has been shown to be NP-hard [71], a number of approximations exist [72], and a comparison by Snarey *et al.* [73] suggests that the $MaxMin$ algorithm of Polinsky *et al.* [74] selects the most diverse subset. The $MaxMin$ algorithm is given in Algorithm 2.1, and the top 15 results when seeded with the GM15510 cell line are given in Table 2.1.

Cell Culture and DNA Preparation

As the acquisition and culture of these cell lines represents the creation of a significant laboratory resource for future studies, I describe their creation in significant detail. Epstein-Barr virus-immortalized lymphoblast cell lines GM15510, GM10860, GM18994, GM18505, GM11839, GM18501 and GM19202 were purchased from the Coriell Cell Repositories, Coriell Institute for Medical Science, Camden NJ. Cells were

cultured in RPMI-1640 with 2 mM L-glutamine (Invitrogen) supplemented with 15% uninactivated fetal bovine serum (Invitrogen) and incubated at 37 °C , 85% relative humidity, 5% CO_2. They were grown in 25 ml cell culture flasks and split twice, at 5 and 10 days, at a ratio of 1:5. After 15 days, 12 flasks of cells were available; assuming a plateau concentration of approximately 1 million cells/ml, this culture method resulted in 300 million cells per cell line.

Cells were harvested in 3 different manners. The first was freezing whole cells: 150 million cells, half the total culture volume, were frozen in freezing medium consisting of RPMI-1640 with L-glutamine, 15% uninactivated fetal bovine serum and 6% DMSO (cell-culture grade from the ATCC) as per the Coriell CCR website. They were spun in 50 ml conicals at 100xg, 4 °C for 10 minutes. The supernatant was decanted and the cells resuspended in ice-cold freezing medium at a concentration of 5 million cells/ml. The cells were aliquotted into 1 ml screw-cap conical tubes and placed immediately on ice. They were transferred to a microcentrifuge tube storage box contained in another styrofoam box, which was filled with ice-cold isopropanol, and the whole thing was transferred into the -80 °C freezer overnight. After freezing overnight, the tubes were transferred to dry ice and labeled. Half were stored in the Dove Lab's liquid nitrogen freezer (Box 1, spots 64-69 and 73-81, contact: Cheri Pasch) so that they can be revived if necessary. The other half were stored in our -80 °C freezer for additional DNA preparations.

The second harvest method was the creation of liquid lysates. Two different lysate formulations were prepared, each at 5 different cell concentrations. The first was the standard heat lysis: 1 ml of cells was pelleted at 100xg at 4 °C , resuspended in the same volume of ice-cold phosphate-buffered saline (PBS), pelleted again, and resuspended again in 1 ml of ice-cold PBS. Cells were counted and diluted to 100, 50,

25, 12, and 6 cells/ul, aliquotted out into 10 aliquots of 100 ul each, and then mixed with an equal volume of TE buffer + 1 mg/ml proteinase K. The tubes were heated for 15 minutes in a 70 °C oven, cooled to room temperature, and stored at 4 °C .

The second batch of liquid lysates was prepared in the manner I described earlier in the chapter. Cells were washed and diluted as above, but were mixed 1:1 instead with a solution of 20 mM EDTA pH 8, 1 mg/ml proteinase K, 10 mM N-lauroylsarcosine (sodium salt) and 2 mM Tris buffer pH 8. These were heated in a 50 °C water bath for 2 hours, cooled to room temperature, and stored at 4 °C .

The final harvest was the creation of gel inserts for pulse field gel electrophoresis, as well as other downstream analyses. A solution of 1.2% low-gel temperature agarose (SeaPlaque Agarose, Cambrex Bio Science, Rockland ME) was prepared in phosphate-buffered saline and equilibrated to 42 °C . Cells were pelleted at 100xg, 4 °C , washed in ice-cold PBS, and resuspended in 5 ml PBS at final cell concentrations corresponding to 50 µg/ml DNA, 25 µg/ml , 12 µg/ml and 6 µg/ml . They were mixed 1:1 with the agarose solution, pipetted into the gel insert mold, chilled, then poked out into a 10 ml solution of 0.5 M EDTA pH 9.5, 1% N-lauroylsarcosine (sodium salt), 1 mg/ml proteinase K. The gel inserts were incubated in this digestion solution for 48 hours at 50 °C , then stored at 4 °C .

Optical Mapping Results

Optical Mapping data sets were created using the preceding protocols for GM15510, GM10860 and GM18994. Collection proceeded until aligned map coverage reached 40-fold. Collection statistics are presented in Table 3.2.

	GM15510	GM10860	GM18994
Input Optical Maps	865,759	1,231,212	1,280,041
Input Optical Map Coverage (fold)	139.15	214.18	220.82
Aligned Optical Maps	215,091	243,751	239,406
Aligned Optical Map Coverage (fold)	46.78	42.11	49.70
Surfaces Collected	515	128	293
Collection Days	65	16	37

Table 2.2. Optical Mapping data collection statistics.

2.3 Conclusion

This chapter described the optimization of the protocols involved in collecting Optical Mapping data, achieving a substantial increase in data quality and a corresponding shortening of data collection times. The additional throughput enabled the creation of several Optical Mapping data sets of normal human genomes in order to study genome structure polymorphism across a small population.

Even with increased throughput, we were limited to only a handful of samples for analysis, making careful sample selection important. The genomes for analysis were chosen based on (a) the availability of other analyses with which to compare our results, and (b) the desire to maximize genetic diversity and thus the breadth of our results. This chapter characterizes the successful creation of several Optical Mapping data sets; the next chapter details the computational methods we used to analyze those data sets and presents the results of those analyses, a survey of normal human genome polymorphism as discerned by Optical Mapping.

2.4 Bibliography

[1] Bailey, J. A. and Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics*, **7**(7):552–64, 2006.

[2] Eichler, E. E. Widening the spectrum of human genetic variation. *Nature genetics*, **38**(1):9–11, 2006.

[3] Zogopoulos, G., et al. Germ-line DNA copy number variation frequencies in a large North American population. *Human genetics*, **122**(3-4):345–53, 2007.

[4] Hastings, P. J., et al. Mechanisms of change in gene copy number. *Nature reviews Genetics*, **10**(8):551–64, 2009.

[5] Kidd, J. M., et al. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods*, **7**(5):365–71, 2010.

[6] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature*, **470**(7333):187–97, 2011.

[7] Baker, M. Structural variation: the genome's hidden architecture. *Nature Methods*, **9**(2):133–137, 2012.

[8] Lai, Z., et al. A shotgun optical map of the entire Plasmodium falciparum genome. *Nature genetics*, **23**(3):309–13, 1999.

[9] Jing, J., et al. Optical mapping of Plasmodium falciparum chromosome 2. *Genome research*, **9**(2):175–81, 1999.

[10] Lin, J., et al. Whole-genome shotgun optical mapping of Deinococcus radiodurans. *Science (New York, NY)*, **285**(5433):1558–62, 1999.

[11] Qi, R. *Whole genome shotgun optical mapping of the Deinococcus radiodurans and Trypanosoma brucei genomes : and the development of a new surface system for optical mapping.* Ph.D. thesis, New York University, 1999.

[12] Lim, A., et al. Shotgun optical maps of the whole Escherichia coli O157:H7 genome. *Genome research*, **11**(9):1584–93, 2001.

[13] Zhou, S., et al. A whole-genome shotgun optical map of Yersinia pestis strain KIM. *Applied and environmental microbiology*, **68**(12):6321–31, 2002.

[14] Zhou, S., et al. Whole-genome shotgun optical mapping of Rhodobacter sphaeroides strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome research*, **13**(9):2142–51, 2003.

[15] Lim, S. A. *Single Molecule Systems: Advancements and Applications to Microbial and Human Genome Analysis.* Ph.D. thesis, University of Wisconsin-Madison, 2004.

[16] Machida, M., et al. Genome sequencing and analysis of Aspergillus oryzae. *Nature*, **438**(7071):1157–61, 2005.

[17] Reslewic, S. *The Optical Mapping of Genomes: Gaining New Insignts on Genome Structure and Variation by Single DNA Molecule Analysis.* Ph.D. thesis, The University of Wisconsin – Madison, 2005.

[18] Herschleb, J. *Optical mapping reveals gene rearrangements and high-resolution structural alterations in breast cancer genomes.* Ph.D. thesis, University of Wisconsin – Madison, 2009.

[19] Sarkar, D. *Analyzing Optical Mapping Data using in silico Restriction Maps*. Ph.D. thesis, University of Wisconsin, 2005.

[20] Sarkar, D. *On the Analysis of Optical Mapping Data*. Ph.D. thesis, University of Wisconsin – Madison, 2006.

[21] Valouev, A., et al. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, **22**(10):1217–24, 2006.

[22] Valouev, A., et al. Alignment of optical maps. *Journal of computational biology : a journal of computational molecular cell biology*, **13**(2):442–62, 2006.

[23] Valouev, A. *Shotgun Optical Mapping: A Comprehensive Statistical and Computational Analysis*. Ph.D. thesis, University of Southern California, 2006.

[24] Valouev, A., et al. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(43):15,770–5, 2006.

[25] Church, D. M., et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, **7**(5):e1000,112, 2009.

[26] Teague, B., et al. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24):10,848–53, 2010.

[27] Sarkar, D., et al. Statistical significance of optical map alignments. *Journal of computational biology : a journal of computational molecular cell biology*, **19**(5):478–92, 2012.

[28] Cai, W., et al. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(11):5164–8, 1995.

[29] Reed, J., et al. A quantitative study of optical mapping surfaces by atomic force microscopy and restriction endonuclease digestion assays. *Analytical biochemistry*, **259**(1):80–8, 1998.

[30] Dimalanta, E. T., et al. A microfluidic system for large DNA molecule arrays. *Analytical chemistry*, **76**(18):5293–301, 2004.

[31] Duffy, D. C., et al. Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Analytical chemistry*, **70**(23):4974–84, 1998.

[32] Zimm, B. H. Dynamics of Polymer Molecules in Dilute Solution: Viscoelasticity, Flow Birefringence and Dielectric Loss. *The Journal of Chemical Physics*, **24**(2):269, 1956.

[33] Zimm, B. H. and Kilb, R. W. Dynamics of branched polymer molecules in dilute solution. *Journal of Polymer Science*, **37**(131):19–42, 1959.

[34] Zhou, S., et al. A single molecule scaffold for the maize genome. *PLoS genetics*, **5**(11):e1000,711, 2009.

[35] Birren, B. and Lai, E. *Pulsed Field Gel Electrophoresis: A Practical Guide*. Academic Press, Inc., 1992.

[36] Caligur, V. Detergents and Solubilization Reagents. Technical Report 3, Sigma Life Science, 2008.

[37] Jing, J., et al. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(14):8046–51, 1998.

[38] Aston, C., Hiort, C., and Schwartz, D. C. Optical mapping: an approach for fine mapping. *Methods in enzymology*, **303**:55–73, 1999.

[39] Zhou, S., et al. Shotgun optical mapping of the entire Leishmania major Friedlin genome. *Molecular and biochemical parasitology*, **138**(1):97–106, 2004.

[40] Zhou, S., et al. Validation of rice genome sequence by optical mapping. *BMC genomics*, **8**:278, 2007.

[41] Ananiev, G. E., et al. Optical mapping discerns genome wide DNA methylation profiles. *BMC molecular biology*, **9**:68, 2008.

[42] Bajorath, J., Hinrichs, W., and Saenger, W. The enzymatic activity of proteinase K is controlled by calcium. *European journal of biochemistry / FEBS*, **176**(2):441–7, 1988.

[43] Ebeling, W., et al. Proteinase K from Tritirachium album Limber. *European journal of biochemistry / FEBS*, **47**(1):91–7, 1974.

[44] Cowan, J. A. Metal Activation of Enzymes in Nucleic Acid Biochemistry. *Chemical reviews*, **98**(3):1067–1088, 1998.

[45] Kavenoff, R. and Zimm, B. H. Chromosome-sized DNA molecules from Drosophila. *Chromosoma*, **41**(1):1–27, 1973.

[46] Schwartz, D. C. and Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, **37**(1):67–75, 1984.

[47] Morozov, V. N., et al. New polyacrylamide gel-based methods of sample preparation for optical microscopy: immobilization of DNA molecules for optical mapping. *Journal of microscopy*, **183**(Pt 3):205–14, 1996.

[48] Moon, J. H., et al. Formation of Uniform Aminosilane Thin Layers: An Imine Formation To Measure Relative Surface Density of the Amine Group. *Langmuir*, **12**(20):4621–4624, 1996.

[49] Moon, J. H., et al. Absolute Surface Density of the Amine Group of the Aminosilylated Thin Layers: Ultraviolet-Visible Spectroscopy, Second Harmonic Generation, and Synchrotron-Radiation Photoelectron Spectroscopy Study. *Langmuir*, **13**(16):4305–4310, 1997.

[50] Kallury, K. M. R., Macdonald, P. M., and Thompson, M. Effect of Surface Water and Base Catalysis on the Silanization of Silica by (Aminopropyl)alkoxysilanes Studied by X-ray Photoelectron Spectroscopy and 13C Cross-Polarization / Magic Angle Spinning Nuclear Magnetic Resonance. *Langmuir*, (16):492–499, 1994.

[51] Matheron, G. Traité de géostatistique appliquée. *Editions Technip*, 1962.

[52] Krige, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand.". *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**(6):119–139, 1951.

[53] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3):443–53, 1970.

[54] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1):195–7, 1981.

[55] Waterman, M. S., Smith, T. F., and Katcher, H. L. Algorithms for restriction map comparisons. *Nucleic acids research*, **12**(1 Pt 1):237–42, 1984.

[56] Menter, P. Acrylamide Polymerization – A Practical Approach. Technical Report 1156 rev. E, Bio-Rad Laboratories.

[57] Stanton, J., et al. Protection of DNA from high LET radiation by two OH radical scavengers, tris (hydroxymethyl) aminomethane and 2-mercaptoethanol. *Radiation and environmental biophysics*, **32**(1):21–32, 1993.

[58] Stein, L. D. Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011):931–45, 2004.

[59] Tuzun, E., et al. Fine-scale structural variation of the human genome. *Nature genetics*, **37**(7):727–32, 2005.

[60] McCarroll, S. A., et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, **40**(10):1166–74, 2008.

[61] Korbel, J. O., et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, NY)*, **318**(5849):420–6, 2007.

[62] Chen, Z., et al. Linear time probabilistic algorithms for the singular haplotype reconstruction problem from SNP fragments. *Journal of computational biology : a journal of computational molecular cell biology*, **15**(5):535–46, 2008.

[63] Hormozdiari, F., et al. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, **19**(7):1270–8, 2009.

[64] Barbujani, G., et al. An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(9):4516–9, 1997.

[65] Foster, M. W. and Sharp, R. R. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome research*, **12**(6):844–50, 2002.

[66] Hinds, D. A., et al. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature genetics*, **38**(1):82–5, 2006.

[67] Hamming, R. W. Error detecting and error correcting codes. *Bell Syst Tech J*, **29**:147–160, 1950.

[68] The International HapMap Consortium. The International HapMap Project. *Nature*, **426**(6968):789–96, 2003.

[69] Goldstein, D. B. and Cavalleri, G. L. A haplotype map of the human genome. *Nature*, **437**(7063):1299–320, 2005.

[70] McVean, G., Spencer, C. C. A., and Chaix, R. Perspectives on human genetic variation from the HapMap Project. *PLoS genetics*, **1**(4):e54, 2005.

[71] Kuo, C.-C., Glover, F., and Dhir, K. S. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sci*, **24**(6):1171–1185, 1993.

[72] Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of computational biology : a journal of computational molecular cell biology*, **6**(3-4):447–57, 1999.

[73] Snarey, M., et al. Comparison of algorithms for dissimilarity-based compound selection. *Journal of molecular graphics & modelling*, **15**(6):372–85, 1997.

[74] Polinsky, A., et al. Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery. chapter LiBrain: S, pages 219–232. American Chemical Society, 1996.

3 STRUCTURAL VARIATION IN FOUR NORMAL HUMAN GENOMES

Having collected an ensemble of single-molecule restriction maps, we are left with the formidable task of drawing biological meaning from them. For the purposes of this thesis, our goal is to discover variation in the structure of the genomes we analyzed, following from our hypothesis that Optical Mapping's unique freedom from ascertainment bias will provide significant additional power to do so when compared to existing methods.

This chapter details the procedures used to analyze the collection of single-molecule restriction maps generated by the Optical Mapping system. I describe the iterative assembly process by which we infer accurate genome-wide structure from local, error-prone observations and the procedures by which we compare those results to a reference. Additionally, I present a detailed comparison of the genome structure variants discerned by Optical Mapping with those described by other authors using other methods. These comparisons not only validate our results but also highlight the scope, power and sensitivity of the Optical Mapping platform.

3.1 Iterative Consensus Map Assembly

As detailed in the previous chapter, I used the Optical Mapping platform [1–14] to generate shotgun single-molecule restriction maps from the genomes of three lymphoblast-derived cell lines (GM15510, GM10860, GM18994). Also included in the following analysis is an Optical Mapping data set from a complete hydatidiform mole (CHM) [15–17], collected by S. Reslewic [18].

Upon completing data collection, the single-molecule restriction maps are assembled into consensus maps using an iterative process (Figure 3.1) whose overall strategy is similar to the operation of the sequence assembler Phusion [19]. First, the single-molecule maps are aligned to a reference map generated *in silico* from the NCBI Build 35 human genome reference sequence; this serves to cluster together similar maps. The pairwise alignments are accomplished with laboratory-developed software called Software for Optical Mapping Analysis (SOMA) (Scott Kohn, unpublished), which uses a modified Needleman-Wunsch [20] dynamic programming algorithm to find an optimal global alignment between two restriction maps.

Next, the single-molecule restriction maps are assigned to clusters based on the location in the genome to which they aligned, and each cluster of maps is assembled into a consensus restriction map by a Bayesian maximum-likelihood assembler called Gentig [22]. Gentig's operation is similar to that of Phrap [23] in sequence assembly: it produces a consensus restriction map that maximizes the likelihood of having produced the observed single-molecule maps, given a model of Optical Mapping errors (described briefly in Table 3.1 and in greater detail in the next chapter). Because Gentig tends to produce spurious assemblies at the end of its map contigs, we trim the consensus maps' flanks to a depth of 4 single-molecule maps, and discard any contigs that were assembled from fewer than 7 maps. The consensus maps that remain are used in place of the NCBI Build 35 reference maps in the next round of clustering and assembly, iteratively refining and extending the consensus restriction maps. Map assembly continues for 8 iterations, refining and extending the consensus maps into a successively more accurate and comprehensive representation of the genome from which the Optical Mapping data set was generated (Figure 3.2).

The genome-wide consensus maps thus constructed span over 95% of the euchro-

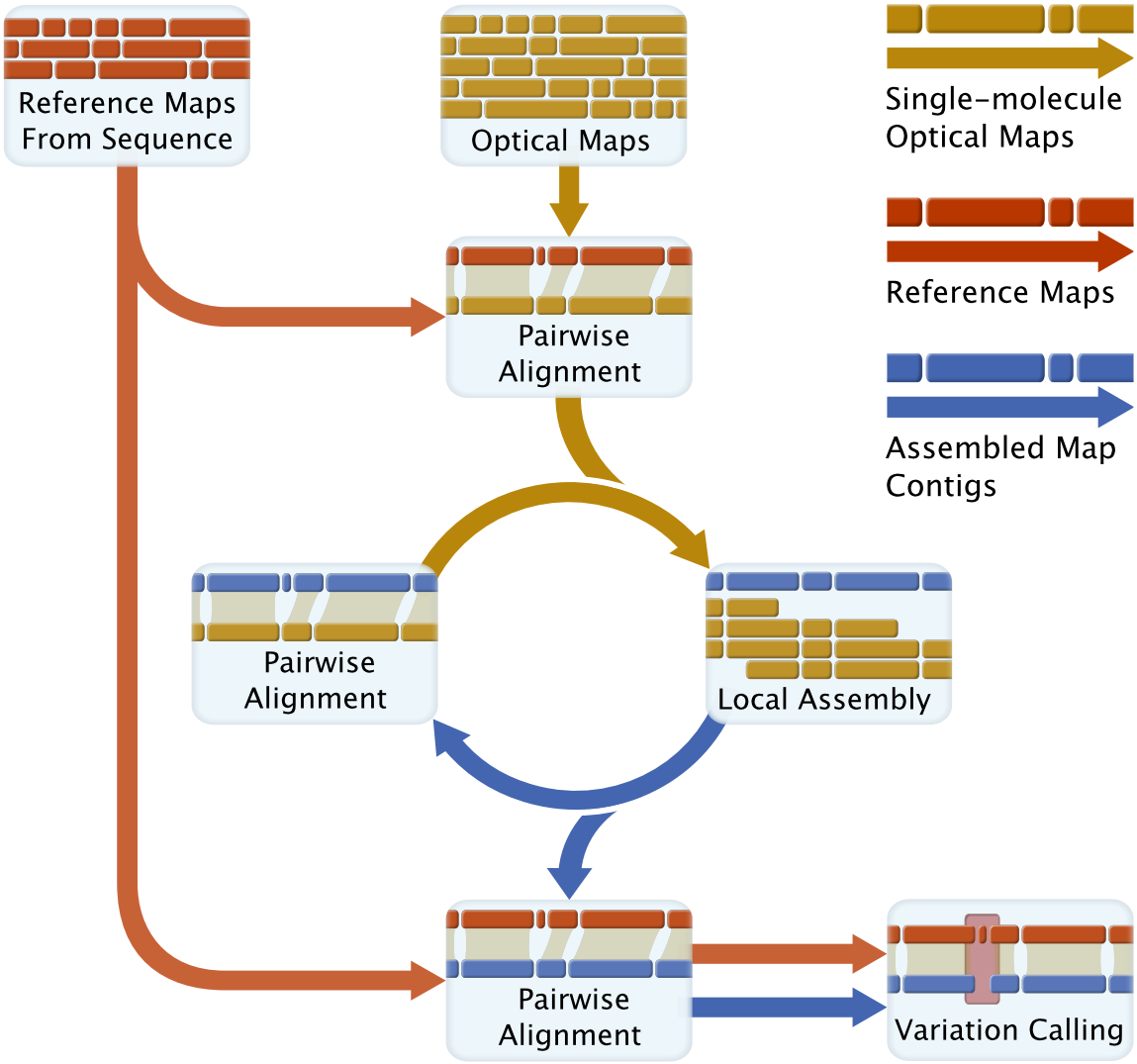

Figure 3.1. An overview of the map assembly pipeline. Reference maps are generated *in silico* from the NCBI Build 35 human genome reference sequence [21], and used to seed an iterative process of pairwise alignment (which clusters together similar single-molecule maps) and local assembly (which generates a consensus optical map from a cluster of single-molecule maps.) After several iterations of alignment and assembly, the consensus maps are aligned back to the reference map and analyzed for places where the consensus map differs significantly from the reference, indicating potential polymorphisms.

Error process	Distribution	Typical parameters
Enzyme doesn't cut at restriction site	$Binomial(p)$	Enzyme cut efficiency $p = 0.8$
DNA is randomly broken	$Poisson(\lambda)$	False cut rate $\lambda = 0.005$ cut/kb
Staining is uneven	$Normal(\mu, V(\mu, \mu^2))$	Actual fragment size $= \mu$
Very small fragments desorb	$Exp(ln(2)/x)$	Median missing frag size $x = 1.35$

Table 3.1. Optical Mapping error processes

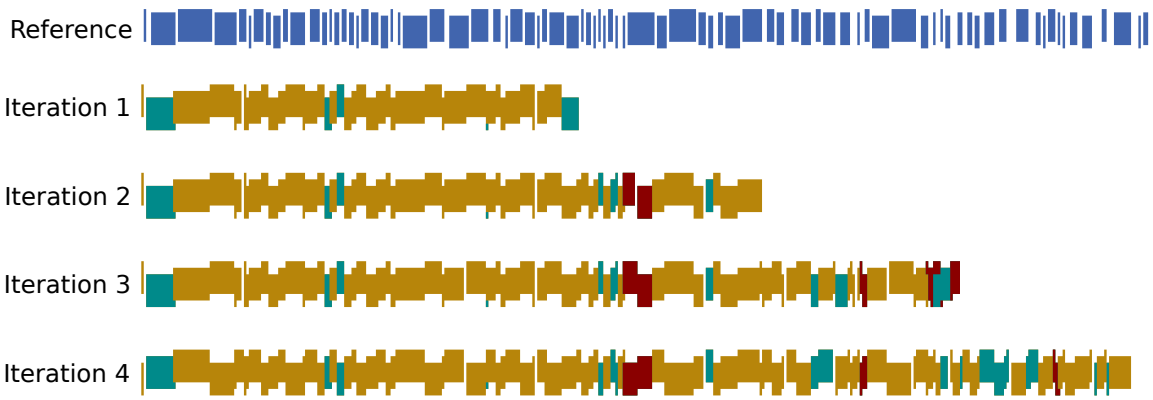

Figure 3.2. Iterative assembly extends and refines consensus maps. As iterative assembly proceeds, consensus maps grow and become a more faithful representation of the genome being studied.

matic genome, and they have an average assembly depth of about 23-fold (CHM) and 46 to 65-fold (GM15510, GM10860, GM18994) (Table 3.2). Because of the iterative nature of the assembly pipeline, consensus maps are not confined to finished sequence and frequently span gaps in the reference sequence; of the 279 gaps in the Build 35 sequence, 170 are spanned by at least one assembly, with 164 having reliable size estimates.

The genome-wide consensus maps are also highly accurate: in all four genomes,

	CHM	GM15510	GM10860	GM18994
Input Optical Maps	416,284	865,759	1,231,212	1,280,041
Input Optical Map Coverage (fold)	65.91	139.15	214.18	220.82
Assembled Optical Maps	110,344	237,012	275,198	301,584
Assembled Optical Map Coverage (fold)	23.51	46.90	60.87	65.25
Consensus Maps	671	2,915	3,352	7,931
Average Consensus Map Size (kb)	4,094	3,139	3,134	2,574
Sequence Scaffold Coverage (%)	96.29	97.36	98.62	98.29

Table 3.2. Optical map collection and assembly statistics.

over 95% of the fragments size 10 kb and greater are within 10% of their corresponding reference fragment size. (This regime's fragment sizing error increases substantially with fragments smaller than 10 kb; see Figure 3.3.) The accuracy of the consensus maps allows us to compare them with a reference restriction map generated *in silico* from a reference sequence and discern differences between the two with high confidence. Protocol 3.1 outlines the rules used to automatically identify differences between the consensus maps and the reference map; the automated variation caller is followed by a manual curation.

3.2 Four Thousand Structural Variants from Four Genomes

Structural Variation Discernment

In order to identify sites of structural variation, we compared the consensus restriction maps to a restriction map generated *in silico* from the NCBI build 35 human genome reference sequence [21] We summarize our findings in Figure 3.4 and Table 3.3. The

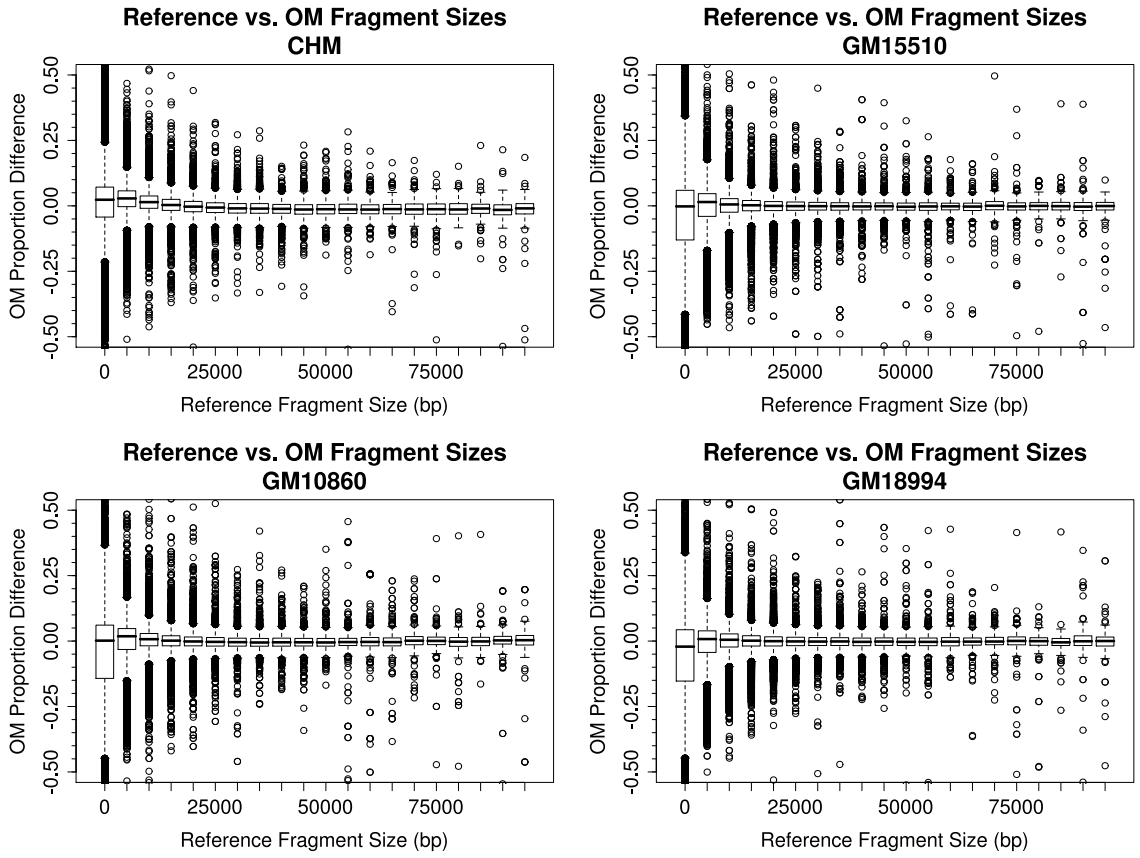

Figure 3.3. The Optical Map is highly accurate. Boxplots represent the proportional difference between the consensus map's fragments and the reference fragments to which they are aligned, broken into 5 kb bins by reference fragment size. The spread of the distribution is substantially larger for fragments 10 kb and smaller than it is for the rest of the genome. The remainder of the map is highly accurate, with over 95% of the consensus fragments having sizes within 5% of their respective reference fragments.

Protocol 3.1 Rules for automated variation calling

The error processes inherent in Optical Mapping make it unlikely that a single molecule map would exactly match the map arrived at if the molecule's entire sequence were known. However, the high-throughput nature of the Optical Mapping platform means that many observations are available at any particular locus; in this context, the error model above provides a structure with which to estimate the statistical significance of a given difference between a consensus optical map and an in silico map based on the reference sequence.

- Missing cuts and extra cuts are assigned a p value based on a binomial test with the parameters given in Table 3.1.
- Insertions and deletions are assigned a p value using a Z-test with the parameters described in Table 3.1.
- Each variant's p value is converted to a false discovery rate-corrected q value using the method of Benjamini and Hochberg [24].
- Our default analysis rejects variants with a q value of less that 0.1. Additionally, we apply several empirically determined filters to the automated analysis, to account for other sources of error from both the experimental and analytical components of our platform:
  - We discard extra cuts and missing cuts if they occur in, or are flanked by, fragments of size less than 4 kb. This accounts for errors induced by the small fragment desorption.
  - We don't consider variants if they are supported by fewer than 10 optical maps in an assembly. Because the CHM assembly has an average depth of about 24 and the lymphoblast-derived cell line assemblies have average depths from about 45 to over 60, we think it likely that such shallow regions likely represent problematic assemblies.
  - We don't consider indels that have reference or consensus fragments smaller than 5 kb. As demonstrated in Figure 3.3, the reliability of our sizing goes down substantially for fragments that small.
  - We don't consider insertions that flank fragments smaller than 5 kb. Gentig has a tendency to merge small fragments with large ones, creating false insertions from the combination of a large fragment and a small one.
  - We discard indels with an absolute size difference of less than 3 kb. This combats the false positives introduced by Gentig's tendency to over-size large fragments.

	Variant type					Variant intersections				
	EC	MC	Ins	Del	Other	Unique	Int.1	Int.2	Int.3	Total
CHM	465	446	165	183	96	471	283	273	322	1355
GM15510	556	384	447	105	105	616	387	417	322	1753
GM10860	584	352	631	350	86	777	447	411	322	2003
GM18994	535	409	523	384	90	735	443	411	322	1941
Total	2140	1555	1766	1214	377	2599	780	504	322	4205

Table 3.3. Summary of structural variants discerned by Optical Mapping
EC, extra cut; MC, missing cut; Ins, insertion; Del, deletion; Int.1, variant intersects with a variant on one other map; Int.2, intersects with 2 other maps; Int.3, intersects with all three other maps.

variants include simple events such as extra restriction sites, missing restriction sites, and insertions and deletions with size differences ranging from megabases down to 3 kb. They also include more complex events such as inversions and large discordant regions. In total, we discerned 4,205 unique variants in the four genomes we analyzed. We note that the smaller total from CHM is likely due to the reduced number of single-molecule restriction maps collected from the limited sample available, resulting in a loss of statistical power. We hypothesize, however, that this lower power is somewhat compensated for by the fact that a complete hydatidiform mole is effectively monoploid [16], eliminating the effect of diploidy on an assembler that wasn't designed to accommodate mixed haplotypes.

We also intersected the variants from each genome with variants of the same type from the other three genomes (Table 3.3). We note that over a third of the variants we report were observed in more than one genome, giving us confidence that these results are due to polymorphism and not the spurious result of cell culture artifacts or other random processes. (The infrequency of culture-induced artifacts is also supported by analyses of the HapMap parent-progeny trios [25]). We suggest

Figure 3.4. Structural variation found in four genomes. Chromosomes are shown by the vertical yellow bars, and structural variants found by Optical Mapping are represented by the colored tick marks. Variants from the CHM genome are depicted in green; GM15510 in blue; GM10860 in red and GM18994 in gray. The inset depicts five example differences from the genome of GM10860: an extra cut, a missing cut, a 250 kb deletion, a 150 kb insertion, and a 150 kb inversion.

that the 322 variants common to all four genomes might be due to assembly errors in the NCBI build 35 reference sequence, or they might represent polymorphisms for which the reference sequence reports a minor frequency allele.

Comparison To Other Platforms

We compared our results to those reported by other investigators who applied different technologies to the same samples we analyzed with Optical Mapping (Table 3.4). These analyses serve both to validate the variants we report, as well as to explore the relative power of the Optical Mapping platform to discern genome structure as compared to these other methods.

Fosmid End-Sequencing

The technology with capabilities most closely matching ours is fosmid end-sequencing (FES); we compared our analysis of the GM15510 lymphoblast-derived cell line with results reported by Tuzun *et al.* [26]. Out of the 297 intermediate structural variants (ISVs) that were found by Tuzun *et al.*, we determined that 203 would be detectable by Optical Mapping. (Undetectable ISVs might include, for example, inversions that were contained entirely within a single SwaI restriction fragment.) Of the 203 detectable ISVs, we found supporting optical map evidence for 101, or 50%. A detailed example comparing two ISVs to the corresponding optical mapping evidence is presented in Figure 3.5.

We further compared the sizes of FES-discerned insertions and deletions to those observed with Optical Mapping. In order to increase the likelihood that the findings from each data set are reporting the same sequence-level event, we only included

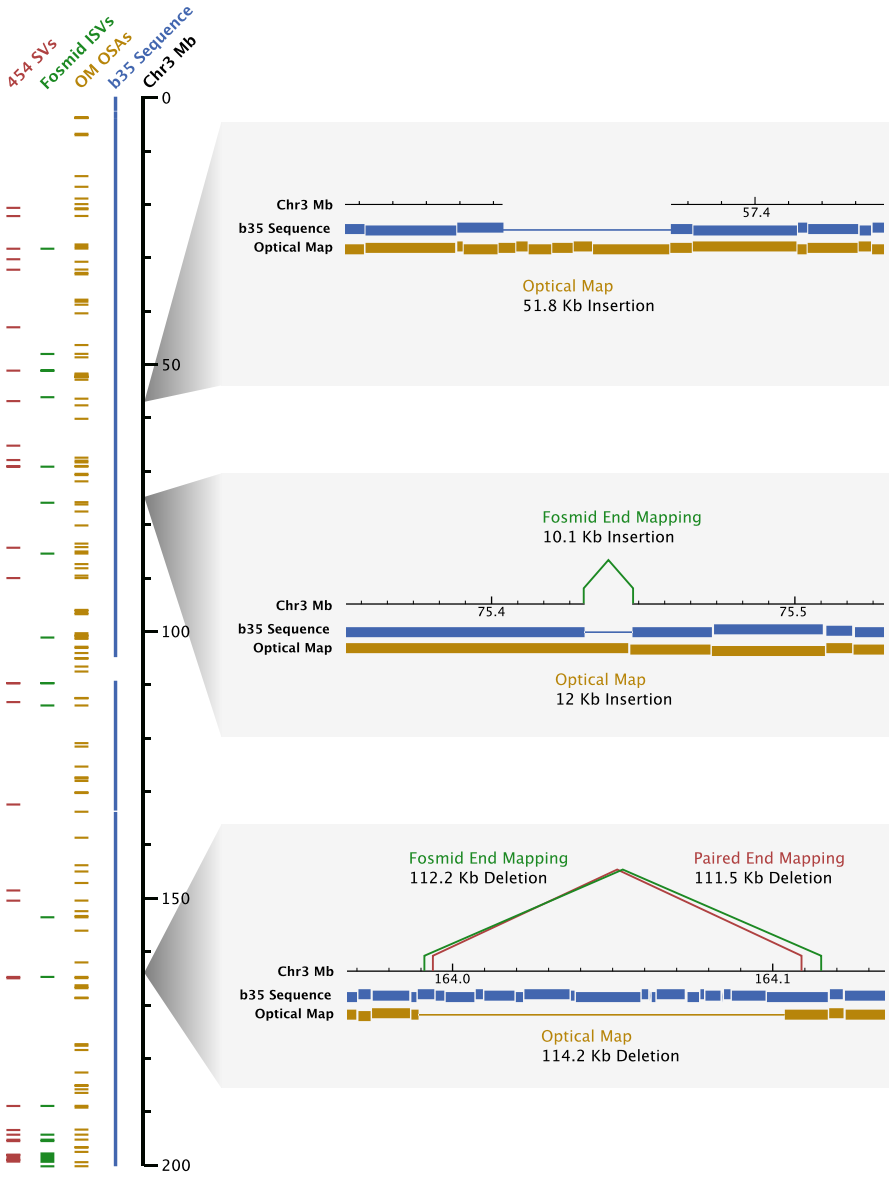

Figure 3.5. A comparison of Optical Mapping results with end mapping methods. The locations of three variants discerned by paired-end mapping, fosmid end-sequencing and Optical Mapping on GM15510's chromosome 3 are presented on the left side of the figure. On the right are three comparisons of the three methods. The first example presents a 52 kb insertion in the optical map; because of their limitations in detecting large insertions, neither sequencing-based technique discovers this insertion. The second shows a smaller 12 kb insertion: the fosmid-based method and the optical map both discern it, but the sequencing-by-synthesis approach does not. The final example presents a 115 kb deletion: all three methods accurately discern it.

Query Platform	Reference Platform: Fosmid End-Sequencing	Paired-End Mapping	Affymetrix SNP 6.0	Tiling Array CGH	Optical Mapping
Fosmid End-Sequencing		92/196 (47%)	262/564 (46%)	262/564 (46%)	58/141 (41%)
Paired-End Mapping	62/109 (57%)		146/163 (90%)	461/641 (72%)	114/473 (24%)
Affymetrix SNP 6.0	562/9527 (6%)	173/753 (23%)		17628/217344 (8%)	93/314 (30%)
Tiling Array CGH	686/9527 (7%)	631/826 (76%)	17628/217344 (8%)		127/1599 (8%)
Optical Mapping	**108/206 (52%)**	**96/231 (42%)**	**33/54 (61%)**	**127/247 (51%)**	

Table 3.4. Summary of Optical Mapping results compared to other platforms A comparison of structural variant detection overlap between several technological platforms when applied to the same samples. Each cell shows the number of variants from the reference platform's results that were detected by the query platform. The reference platform's variants are first filtered to remove those that the query technology is not expected to be able to detect. Fosmid end-sequencing data from [26] and [27]; paired-end mapping data from [28]; Affymetrix CNV data from [25]; tiling array CGH data from [29].

Optical Mapping results that matched one-to-one with an FES-derived observation. We were left with 38 pairs of observations, four of which were discarded after manual curation. A linear model fit to the remaining pairs has an R^2 of 0.98 and a slope of 0.99, indicating strong agreement between the two methods on this set of variants (Figure 3.6 and 3.7).

Finally, additional FES analysis reported by Kidd *et al.* [27] demonstrated that clusters of fosmids with only one aligned end ("OEA fosmids") frequently indicate the presence of an insertion that was too large to be captured by a fosmid library

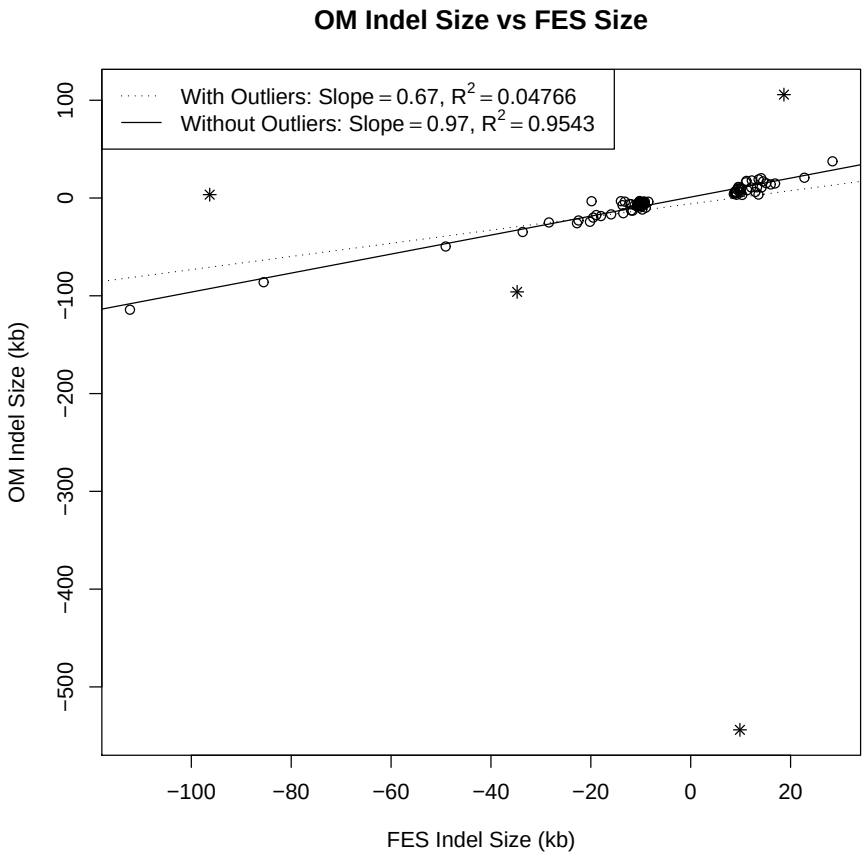

Figure 3.6. Optical Mapping's results show strong concordance with those of fosmid-end sequencing. The plot presents the indel sizes predicted by Optical Mapping and fosmid end-sequencing for loci where there was a one-to-one correspondence between the two methods. Four outliers (shown by stars) were investigated manually, and in each case it was clear that the two methods were not describing the same sequence-level event; an example is given in Figure 3.7. A linear model (solid line) fit to the remaining points has a slope of 0.97 and an R2 of 0.95, indicating strong agreement between the two methods.

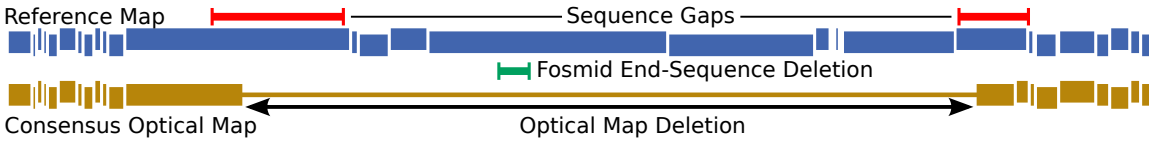

Figure 3.7. One of the outliers discarded from analysis presented in Figure 3.6. The fosmid end-sequencing analysis detects a small deletion (green) between the two gaps in the reference sequence (red); but the Optical Mapping results support a different interpretation, suggesting that the entire region between the sequence gaps is deleted or misassembled.

whose average insert size is 40 kb. Out of 11 clusters identified by Kidd *et al.*, 8 have clear support in the optical map and a ninth comes from a region of large discordance between the optical map and the reference genome, making the presence of extra sequence likely. Two detailed examples, including micrographs of some of the DNA molecules that support this conclusion, are presented in Figure 3.8. We also find evidence that OEA fosmids occurring outside of clusters might indicate smaller insertions: an interval-intersection permutation test (See Protocol 3.2) reveals a significant intersection with optical map-discerned insertions ($P < 0.0001$).

Protocol 3.2 Interval Intersection Permutation Test

1. The singleton fosmids reported by Kidd et al. were filtered to remove clusters of singletons; any singleton within 80 kb of another was removed.
2. The expected interval of each singleton fosmid's read mate was determined based on the mean and standard deviation of the fosmid collection's insert size. An interval width of 3 times the insert size standard deviation was used.
3. The intervals thus constructed were mapped randomly onto the genome; a mapping was accepted only if the intervals were all non-overlapping. This process was repeated until 10,000 accepted mappings were generated.
4. The p value of the test was the proportion of random mappings that intersected the same or greater number of Optical Mapping insertions as the observed result. The test was significant at the level $p = 0.0001$.

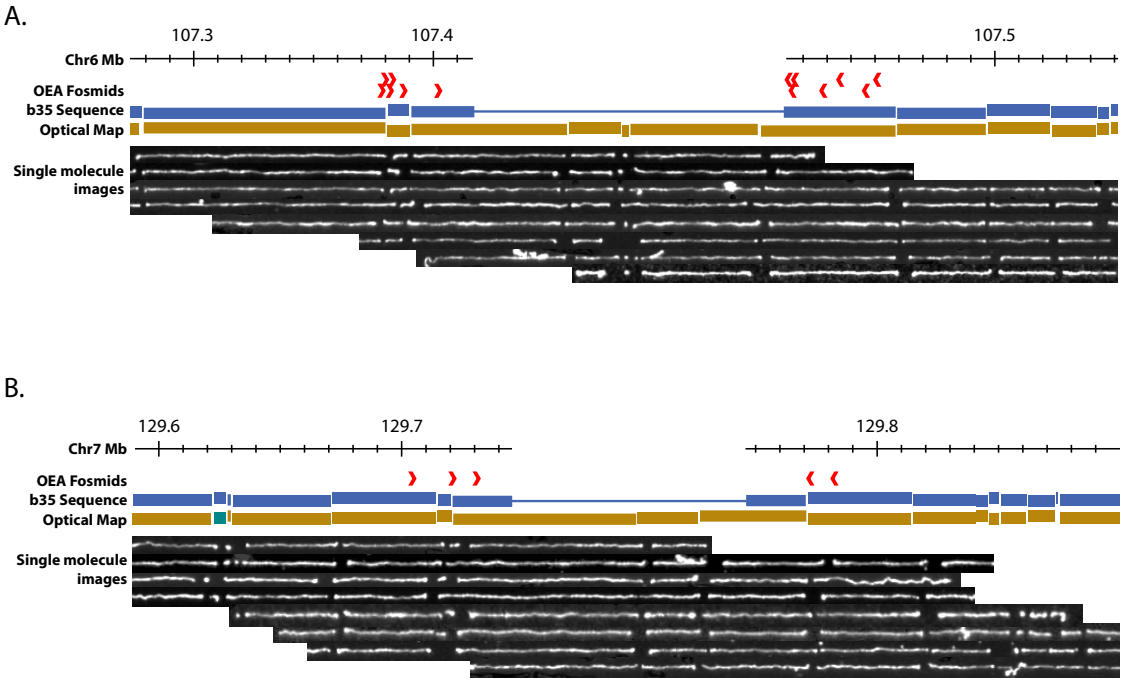

Figure 3.8. Optical Mapping finds large insertions that fosmid end-mapping misses. Two large insertions are shown from GM15510, one from chromosome 6 and one from chromosome 7. Optical Mapping indicates large insertions at both loci, confirming the large insertion that was indicated by a cluster of singleton fosmids reported by Kidd et al. [27] (red arrows). Included below both maps are montages of several of the single-molecule images that give evidence to support this insertion.

Paired-End Sequencing By Synthesis

The advent of next-generation sequencing platforms has extended the hunt for structural variation into the realm of sequencing by synthesis. Korbel *et al.* [28] applied paired-end mapping (PEM) to discover structural variants (SVs) in the GM15510 genome, and we compared our results to theirs as well (Table 3.4). Of the 488 SVs detected by PEM, we determined that 231 would be detectable by Optical Mapping. Of the 231 detectable SVs, we found supporting optical map evidence

for 96. A detailed example comparing an SV to the corresponding optical mapping evidence is presented in Figure 3.5.

We also compared the sizes of PEM-discerned insertions and deletions to those found by Optical Mapping. As with the ISV comparison, we only compared optical mapping results that matched one-to-one with a PEM finding, of which there were 82. We discarded four outliers, three of which overlapped gaps in the reference sequence and one of which gave evidence of a haplotype difference. A linear model fit to the remaining 78 pairs had an R^2 of 0.94 and a slope of 0.98, indicating strong agreement between the two methods on this set of variants (Figure 3.9).

Microarray Platforms

Since the first reports of copy number variations (CNVs) in humans [30], the hybridization of genomic DNA to microarrays has been the most widely employed approach for discovering structural variants. Redon *et al.* [31] applied this technology to the entire HapMap panel (of which GM10860 and GM18994 are members), using both a whole-genome tiling path clone array and an early-access Affymetrix 500K SNP chip. They identified 123 variants using the whole-genome tiling path array, of which 36 (29%) should have been discernable *via* Optical Mapping, while 9 (25%) actually were. The Affymetrix 500K SNP arrays yielded 52 variants from the two HapMap samples, of which 26 (50%) should have been seen with Optical Mapping, while 5 (19%) actually were.

We also performed microarray hybridization experiments on the GM15510 and GM18994 samples using the Affymetrix SNP 6.0 microarray platform; data on GM10860 was available through Affymetrix as part of their HapMap Data Set. Segmentation analysis with the Affymetrix-supplied Genotyping Console 2.1 revealed

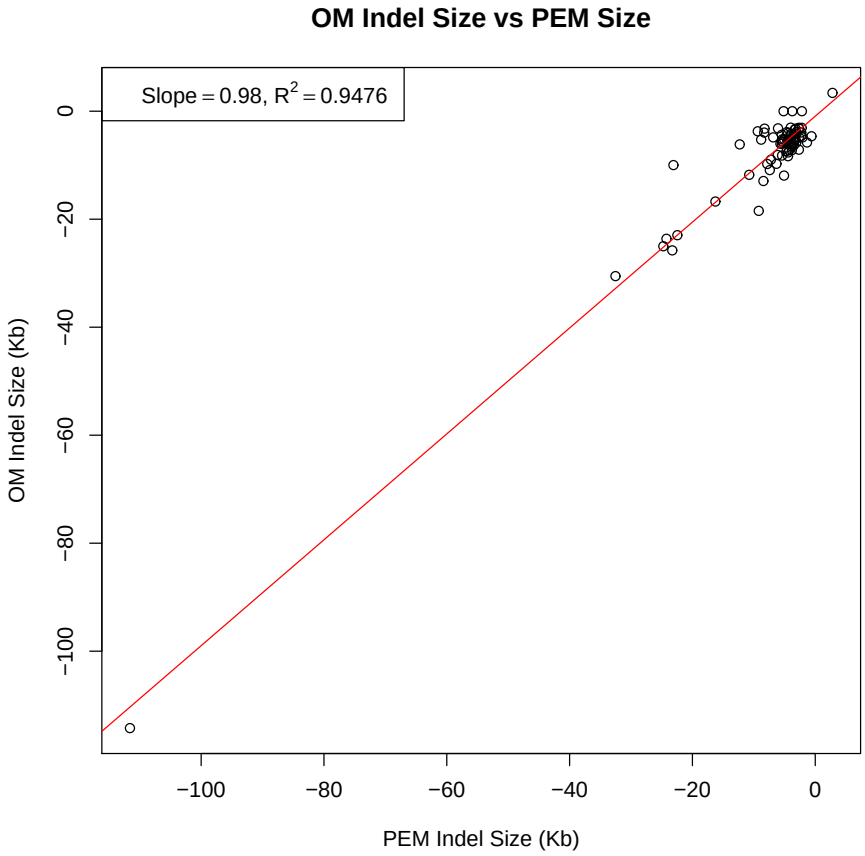

Figure 3.9. Optical Mapping's results show a strong concordance with those of paired-end mapping. The plot presents the indel sizes predicted by Optical Mapping and paired-end mapping for loci where there was a one-to-one correspondence between the two methods. A linear model fit to the pairs of observations has a slope of 0.98 and an R2 of 0.95, indicating a strong agreement between the two methods. Interestingly, only a single insertion was large enough to be passed by our minimum-size filters, but small enough to be detectable by paired-end mapping.

32 changes in copy number, 5 of which have direct support in the optical map data. (Six more come from regions that show major discordance between the optical map and the NCBI reference sequence.) Detailed examples of two events (an insertion and a deletion) are presented in Figure 3.10.

Finally, we compared the SNP genotype calls from the Affymetrix SNP 6.0 platform to the optical map at loci where a SNP creates or ablates a SwaI restriction endonuclease cognate site (i.e. a "snip-SNP." [32]) Two hundred thirty eight such sites exist between the three genomes; two thirds were in regions that were amenable to Optical Mapping analysis. (A snip-SNP that is very close to another cognate site, for example, would be unlikely to show up in our analysis.) Of the 159 amenable snip-SNPs, the Optical Map has strong evidence for 149, a success rate of 94%.

Non-SNP ECs and MCs Are Likely Small Indels

We hypothesized that the extra cuts and missing cuts that did not correspond to snip-SNPs might in fact be the result of insertions and deletions whose sizes were below our ability to resolve. To test this, we segregated ECs and MCs into those that were associated with a snip-SNP and those that weren't, then compared the distributions of the differences between consensus and reference map fragment sizes at these loci. For all three lymphoblast-derived genomes, the consensus fragments with extra cuts due to snip-SNPs are smaller than those with extra cuts without snip-SNPs; and consensus fragments with missing cuts due to snip-SNPs are larger than those with missing cuts but without an associated snip-SNP (Figure 3.11). These differences are highly significant using a Mann-Whitney U test [33], $p < 0.001$ (except for GM15510 extra cuts, $p = 0.016$). These results indicate that, on a global scale, extra cuts and missing cuts that are not associated with snip-SNPs are likely

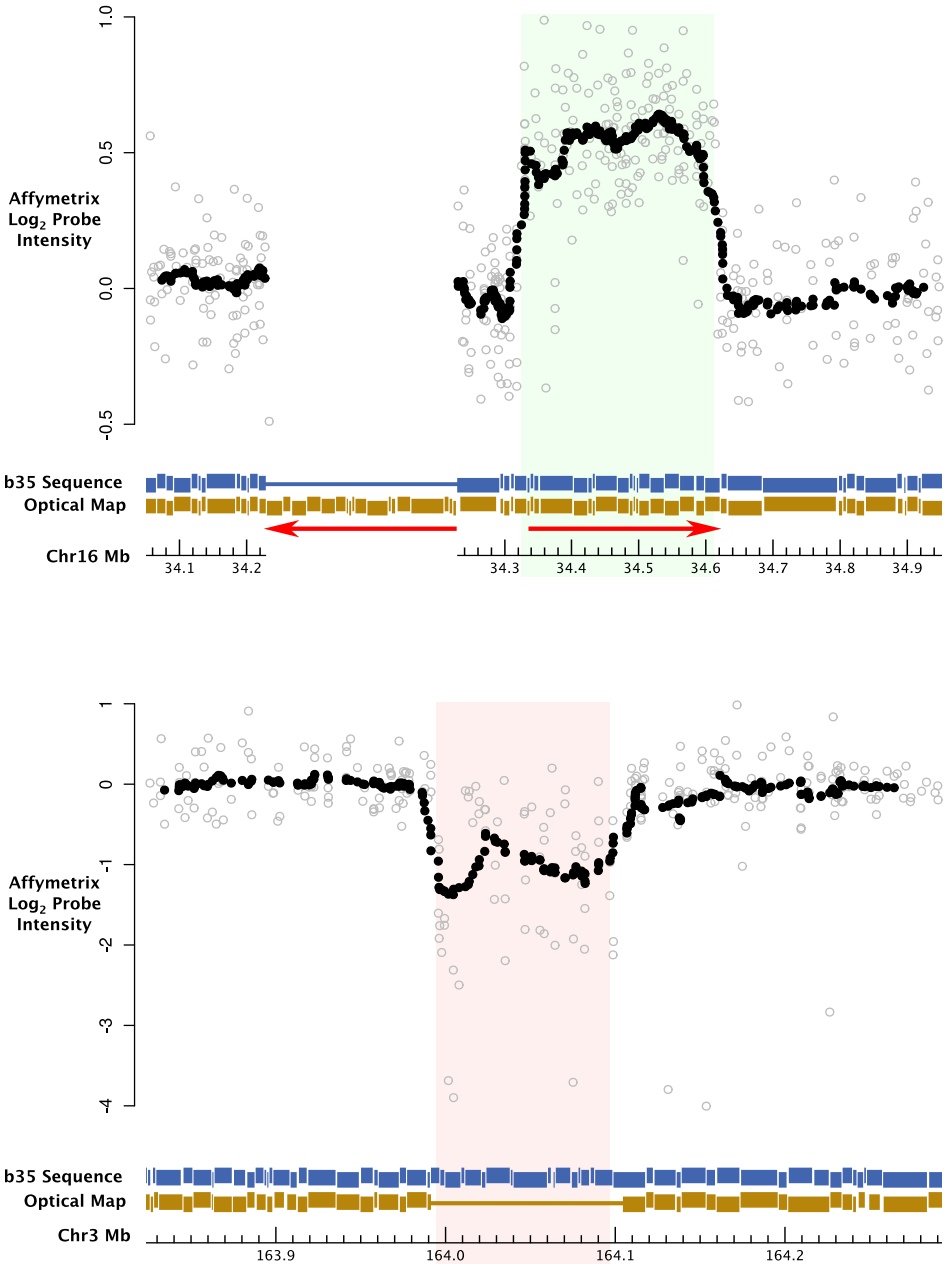

Figure 3.10. The optical map complements hybridization-based approaches. Two examples compare results from the Affymetrix SNP 6.0 platform and Optical Mapping: the first is a 290 kb insertion from GM10860 chromosome 16, and the second is a 114 kb deletion from GM15510 chromosome 3. The optical map confirms both results, and demonstrates that the gain of sequence in the first example is due to an inverted tandem duplication at this locus.

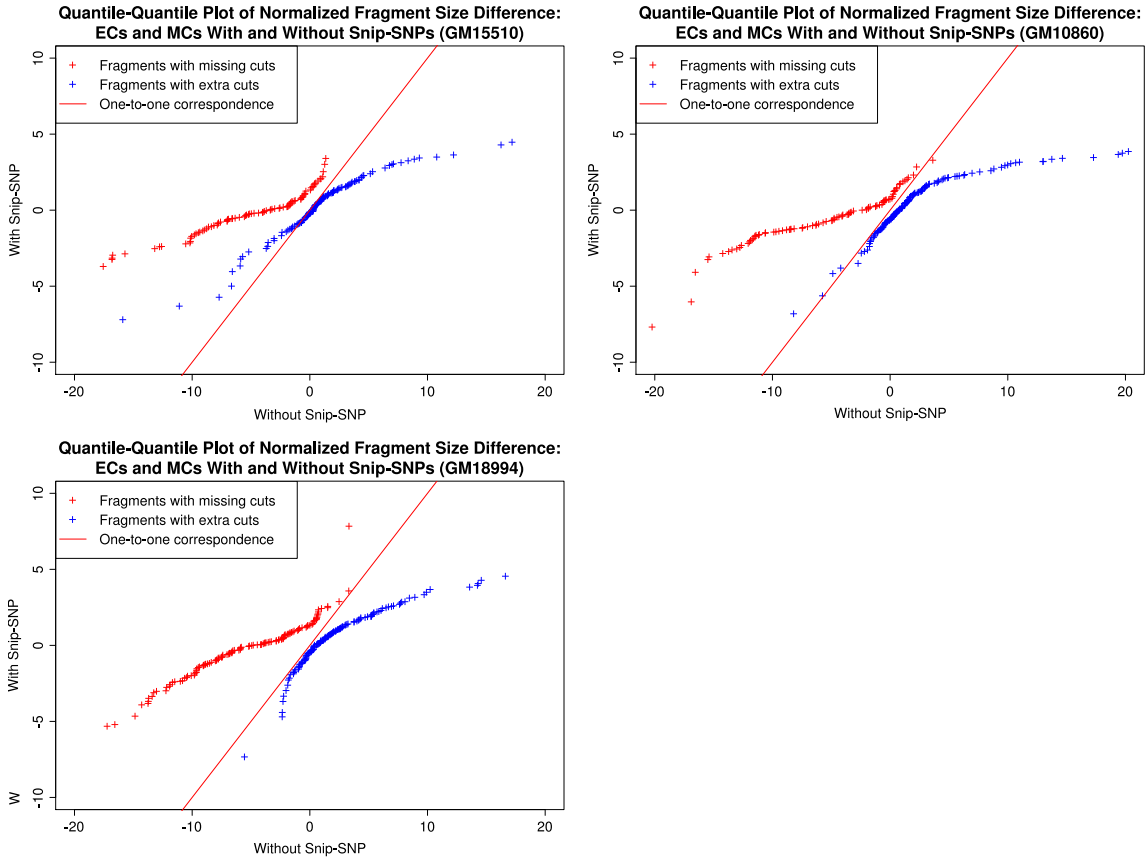

Figure 3.11. Non-SNP extra cuts and missing cuts are likely small indels. Consensus map fragments with extra due to snip-SNPs are smaller than those with extra cuts but without a snip-SNP; and consensus map fragments with missing cuts due to snip-SNPs are larger than those without snip-SNPs. These differences seem to indicate that many of the extra cuts might in fact be small insertions, and many of missing cuts might in fact be small deletions.

to be the result of insertions and deletions below the resolution of our optical map.

Optical Mapping Reveals Variants Inaccessible to Other Platforms

We wanted to determine if Optical Mapping's unique properties quantitatively affect the variants it is able to discern. We focused on repeat-rich regions, because repeats

are closely associated with structural variants [34, 35] but can hamper discernment efforts. We examined the performance of Optical Mapping and the two most current technologies, paired-end mapping [28] and tiling array CGH [29], by classifying each variant as detected by Optical Mapping, detected by the alternate technology, or detected by both. We then compared the proportions of these classes from the entire genome with those subsets that intersect the 6 most common classes of repeat from the UCSC Genome Browser's RepeatMasker database [36] (Figure 3.12a). While the proportion of Optical Mapping-discerned results compared to PEM is about the same in repeat-rich regions as in the entire genome, the repeat-intersecting proportion significantly increases when Optical Mapping is compared to the hybridization-based technology (χ^2 test, $p < 10^{-7}$). We interpret this as evidence that Optical Mapping has a similar power to discern variants in repeat-rich regions as PEM, but a greater capacity in this regard than tiling array CGH.

We also compared the distributions of insertion and deletion sizes between Optical Mapping, PEM and tiling-array CGH. (Figure 3.12b). Optical Mapping is the only platform that does not evidence a strong bias towards the detection of deletions, perhaps due to its lack of reliance on a reference sequence either for probe selection or to anchor end-sequences.

3.3 Conclusion

The Optical Mapping results presented here confirm the prevalence of natural structural variation in the human genome. We present evidence for over 4000 unique structural variants from four normal human genomes, with sizes ranging from several thousand to several million base pairs. We present the substantial overlap in the four

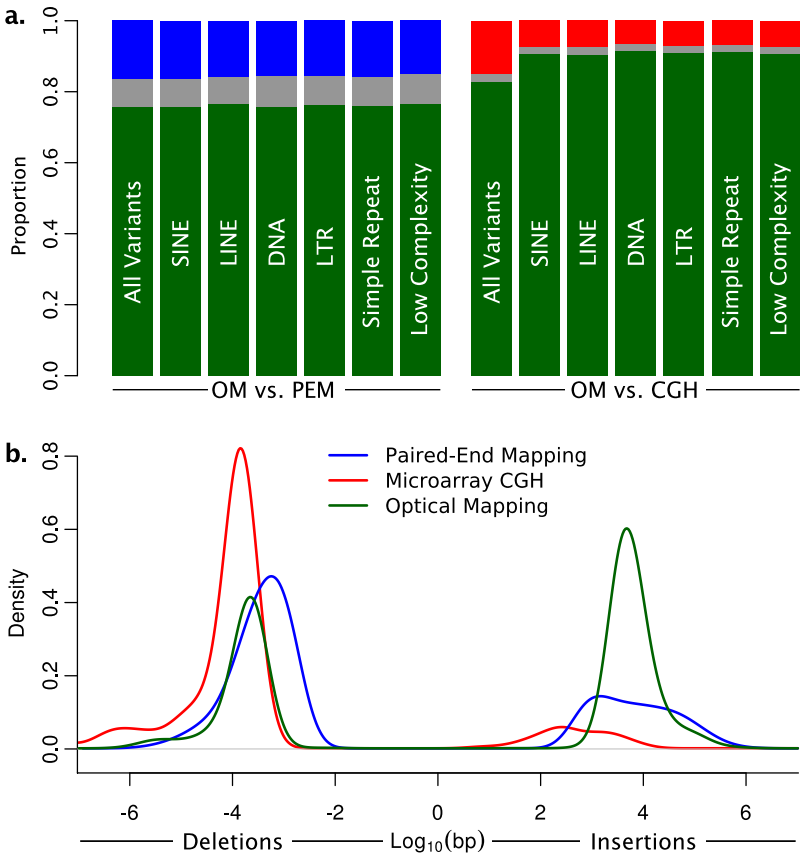

Figure 3.12. Optical Mapping reveals variants inaccessible to other platforms. (a) Optical Mapping has greater ability to discern variation in repeat-rich regions than hybridization-based technologies. The first bar in each section is a genome-wide representation of variants discerned only by Optical Mapping (green), only by an alternate technology (blue for PEM, red for CGH), or by both technologies (gray). (For example, in the first bar of the PEM comparison, 76% of the variants were found only by Optical Mapping, 17% were found only by PEM, and 7% were found by both technologies.) Subsequent bars represent the same proportions, but include only variants that intersect with various classes of repeat. The proportions are substantially the same when comparing Optical Mapping to PEM, but Optical Mapping detects a greater proportion of variants intersecting repeats when compared to hybridization-based technologies (χ^2 test, $p < 10^{-7}$). (b) Optical Mapping-discerned variants are more evenly distributed between insertions (median size, 4.5 kb) and deletions (median size, 4.3 kb). We compared the sizes of indels discovered with Optical Mapping to platforms based on end-sequencing and hybridization. Indel size density was estimated for each dataset using a Gaussian kernel with a bandwidth of 0.3. Negative sizes represent deletions, while positive sizes are insertions. The Optical Mapping indels are more evenly distributed between insertions and deletions, perhaps due to the platform's unique ability to detect large novel insertions.

sets of variants as evidence that the variations we detect are not random experimental error, but instead represent actual sequence-level differences between the analyzed genomes and the NCBI build 35 reference sequence. We support this assertion with discrete observations representing single molecules of DNA from the genomes under study. And we propose that the substantial number of unique variants discerned in just four individuals suggests many additional variants remain undiscovered in the human population as a whole.

We also show that these results confirm and complement the results of other technologies. We show a close concordance with both fosmid end-sequencing [26, 27] and paired-end mapping [28], though the Optical Mapping results are not limited to the small insertions available to these mapping methods. The Optical Mapping results also bring structural insight to insertions and deletions discovered by hybridization-based methods, and are not limited to regions of the genome amenable to unique probe design. These advantages lead to a more balanced distribution of insertions and deletions, an indication of Optical Mapping's low systematic ascertainment bias and its ability to reveal structural variants inaccessible to other platforms. We also note Optical Mapping's ability to handle balanced events such as inversions and rearrangements, areas of genome structural variation that other high-throughput methods are just beginning to explore.

The work presented in this chapter also highlights some of the weaknesses of our current analysis pipeline. For one thing, the variants we discovered comprise only a small portion of the reference genome; most of the computational effort involved in iterative assembly is spent recapitulating genome structure that has not changed. Perhaps a different analysis regime might exploit this fact to circumvent some of this wasteful computation.

Another stumbling block is the discernment of haplotypes. When analyzing human genomes (and absent special circumstances such as the complete hydatidiform mole's effective monoploidy), one is actually analyzing two genomes at once, a situation for which our assembler was not designed. The reporting of haplotypes in our current analysis is haphazard and ad-hoc; perhaps a new analytical framework would provide the basis to approach this problem in a theoretically sound fashion.

The remainder of this thesis describes just such a new framework for the analysis of Optical Mapping data. It is built around a hidden Markov model and provides the basis for addressing the shortcomings in the present analysis.

3.4 Bibliography

[1] Anantharaman, T. S. and Mishra, B. Genomics via Optical Mapping I: Probabilistic Analysis of Optical Mapping Models. Technical Report TR1998-770, New York University, 1998.

[2] Anantharaman, T. S., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. II: Ordered restriction maps. *Journal of computational biology : a journal of computational molecular cell biology*, **4**(2):91–118, 1997.

[3] Anantharaman, T. S., Mishra, B., and Schwartz, D. C. Genomics via Optical Mapping III: Contiging Genomic DNA and Variations. Technical Report TR1998-760, New York University, 1998.

[4] Anantharaman, T. and Mishra, B. Genomics via Optical Mapping II(A): Restriction Maps from Part Molecules and Variations. Technical Report TR1998-759, New York University, 1998.

[5] Jing, J., et al. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(14):8046–51, 1998.

[6] Aston, C., Hiort, C., and Schwartz, D. C. Optical mapping: an approach for fine mapping. *Methods in enzymology*, **303**:55–73, 1999.

[7] Aston, C., Mishra, B., and Schwartz, D. C. Optical mapping and its potential for large-scale sequencing projects. *Trends in biotechnology*, **17**(7):297–302, 1999.

[8] Dimalanta, E. T., et al. A microfluidic system for large DNA molecule arrays. *Analytical chemistry*, **76**(18):5293–301, 2004.

[9] Valouev, A., et al. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, **22**(10):1217–24, 2006.

[10] Valouev, A., et al. Alignment of optical maps. *Journal of computational biology : a journal of computational molecular cell biology*, **13**(2):442–62, 2006.

[11] Valouev, A. *Shotgun Optical Mapping: A Comprehensive Statistical and Computational Analysis*. Ph.D. thesis, University of Southern California, 2006.

[12] Valouev, A., et al. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(43):15,770–5, 2006.

[13] Zhou, S., Herschleb, J., and Schwartz, D. C. *New high throughput technologies for DNA sequencing and genomics*, pages 266–301. Elsevier, 2007.

[14] Teague, B., et al. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24):10,848–53, 2010.

[15] Kajii, T. and Ohama, K. Androgenetic origin of hydatidiform mole. *Nature*, **268**(5621):633–4, 1977.

[16] Jacobs, P. A., et al. Mechanism of origin of complete hydatidiform moles. *Nature*, **286**(5774):714–6, 1980.

[17] Matsuda, T. and Wake, N. Genetics and molecular markers in gestational trophoblastic disease with special reference to their clinical application. *Best practice & research Clinical obstetrics & gynaecology*, **17**(6):827–36, 2003.

[18] Reslewic, S. *The Optical Mapping of Genomes: Gaining New Insignts on Genome Structure and Variation by Single DNA Molecule Analysis*. Ph.D. thesis, The University of Wisconsin – Madison, 2005.

[19] Mullikin, J. C. and Ning, Z. The phusion assembler. *Genome research*, **13**(1):81–90, 2003.

[20] Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3):443–53, 1970.

[21] Stein, L. D. Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011):931–45, 2004.

[22] Anantharaman, T., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. III: Contiging genomic DNA. In *Proc Int Conf Intell Syst Mol Biol*,

Proceedings 7th Intl. Cnf. on Intelligent Systems for Molecular Biology, pages 18–27. 1999.

[23] Gordon, D., Desmarais, C., and Green, P. Automated finishing with autofinish. *Genome research*, **11**(4):614–25, 2001.

[24] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B …*, **57**(1):289–300, 1995.

[25] McCarroll, S. A., et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, **40**(10):1166–74, 2008.

[26] Tuzun, E., et al. Fine-scale structural variation of the human genome. *Nature genetics*, **37**(7):727–32, 2005.

[27] Kidd, J. M., et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191):56–64, 2008.

[28] Korbel, J. O., et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, NY)*, **318**(5849):420–6, 2007.

[29] Conrad, D. F., et al. Origins and functional impact of copy number variation in the human genome. *Nature*, **464**(7289):704–12, 2010.

[30] Lucito, R., et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome research*, **13**(10):2291–305, 2003.

[31] Redon, R., et al. Global variation in copy number in the human genome. *Nature*, **444**(7118):444–454, 2006.

[32] Wicks, S. R., et al. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nature genetics*, **28**(2):160–4, 2001.

[33] Mann, H. B. and Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, **18**(1):50–60, 1947.

[34] Bailey, J. A. and Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics*, **7**(7):552–64, 2006.

[35] van Ommen, G.-J. B. Frequency of new copy number variation in humans. *Nature genetics*, **37**(4):333–4, 2005.

[36] Rhead, B., et al. The UCSC Genome Browser database: update 2010. *Nucleic acids research*, **38**(Database issue):D613–9, 2010.

4 A HIDDEN MARKOV MODEL FOR OPTICAL MAP ANALYSIS

Much of science concerns itself with building models of physical processes to explain data. Sometimes, our interest is in inferring properties or parameters of the process itself: for example, we might take a database of protein structures and use it to calibrate a model of the intramolecular atomic forces that made a linear polypeptide chain fold into its biologically active 3-dimensional conformation [1]. Other times, we want to use a model to better understand the data we observe; we might use the same model to guide the interpretation of X-ray crystallography data, for example [2]. This interplay between models (even if they're just implicit conceptual models) and data is so central to the advancement of scientific understanding that the selection of an appropriate model for the data can determine whether or not an investigator is successful at extracting meaning from her observations.

The structure of the data can be a crucial factor in choosing a model. For example, data frequently show a statistical dependence between observations that are close together (either spatially or temporally). Suppose we're modeling human speech in order to perform speech recognition: both the individual *sounds* (phonemes) and their *order* are important for conveying meaning [3]. One of the most common ways to model this spatial or temportal dependence is with a Markov chain [4] (Figure 4.1). Markov chains represent the physical process that generated the data as inhabiting a finite collection of states $s_{1..n}$, each of which produces an observed output $o_{1..n}$. The Markov chain model also specifies the probabilities which which the system transitions from one state to the next. Importantly, in a Markov chain the system is "memoryless" - the probability of the transition from state s_i to s_j depends only on

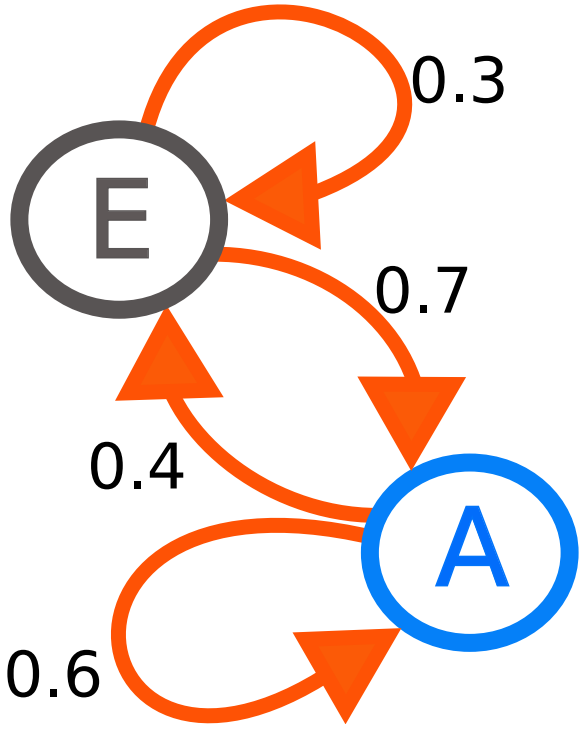

Figure 4.1. A basic Markov chain. Adapted from Wikipedia, http://en.wikipedia.org/wiki/Markov_chain.

the fact that the system is in state s_i and not on any additional history of the system. Such models are particularly useful because writing the transition probabilities as a stochastic matrix lets one use linear algebraic methods to analyze the long-term behavior of the system [5].

However, what if we disconnect the state space from the output space? (Figure 4.2). Now, instead of each state producing one observation, it can produce any output in the output space. We model this by attaching to each state an output distribution, which is a probability distribution over the output space. What this implies is that we can't look at a sequence of observations and tell what sequence of states produced it. This "curtain" between the states the system is in and the sequence of observations is what turns a Markov chain model into a hidden Markov model (HMM) [6].

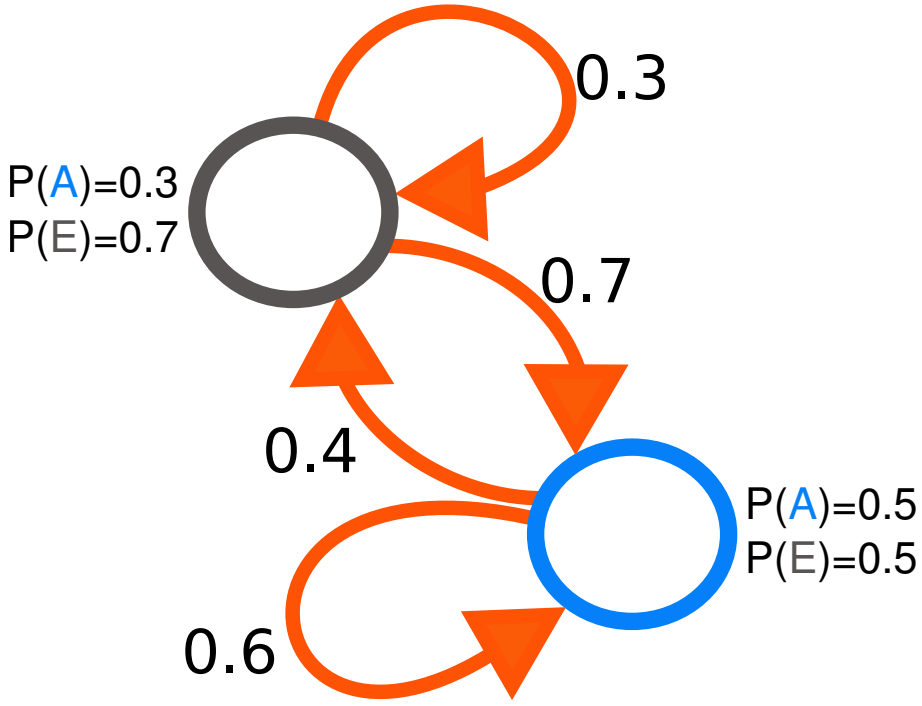

Figure 4.2. A basic hidden Markov model. Adapted from Wikipedia, http://en.wikipedia.org/wiki/Markov_chain.

The power of using a hidden Markov model comes from a corpus of theory that lets us re-establish the link between the state space and the output space [6]. This body of theory (and an accompanying set of efficient algorithms) lets us perform the same kinds of inference discussed above. We can infer things about the model (and thus about the underlying physical process that produced the data), considering questions such as "What transition and emission probabilities maximize the probability that the model produced the data that were observed?" We can also use a given model to perform inference on the data, asking questions like "What is the probability that this model produced a particular output?" and "What sequence of states most likely produced it?"

Because many problems in biological data analysis can be posed in such a way, the

last two decades have witnessed an explosion of HMM-based bioinformatic analyses. Hidden Markov models have been used to create genetic linkage maps [7], mine biological text repositories [8], reconstruct phylogenies [9] and analyze microarrays for copy number changes [10–12]. Their use in nucleotide sequence analysis is manifold: sequence alignment [13], gene finding [14], methylation site detection [15], and splice site prediction [16] have all benefitted. Protein structure analysis has also been a productive application, from secondary structure prediction [17, 18] to structural homology detection [19]. In all these cases, the relevant problems can be cast as a system transitioning between a discrete number of states, producing observable data whose information content is partly or mostly determined by their context.

This introduction to hidden Markov models also makes it clear why HMMs are a good fit for analyzing Optical Mapping data. A single restriction fragment provides a little data about where in the genome it might have come from; but a sequence of those fragments (i.e. an ordered restriction map) provides just the statistical dependence (context) that gives an HMM its inferential power. Optical Mapping also has well-characterized error processes, providing well-founded ways to set transition and emission probabilities based on traditional probability distributions.

Given this apparent fit, it is perhaps no surprise that the work described below is not the first effort to apply a hidden Markov model-based analysis to answer Optical Mapping questions. In 2006, Valouev et al. published [20] a hidden Markov analysis with the aim of iteratively refining reference maps: the algorithm they described started with a model derived from a reference sequence, then iteratively modified the model so as to better represent the genome that produced the single-molecule maps they were analyzing. Though the model they constructed is similar to the one I will describe below, their approach was different in two key ways. First, Valouev

used pairwise alignments as a proxy for posterior decoding, producing a speedier analysis but sacrificing power in doing so. Second, the steps by which they updated their model were informed by a frequentist viewpoint, using traditional statistical tests to choose changes to the model that rejected their null hypothesis, i.e. that the differences between the reference and the single-molecule maps were due to the error processes and not due to a change in the underlying genome. In contrast, I approached the problem from a more classically Bayesian perspective, both to provide more discriminitive power and to expand the range of analyses that can be done within the HMM's theoretical framework.

4.1 A Hidden Markov Model for Optical Mapping

Error Processes in Optical Mapping

The hidden Markov model described in this chapter is derived from models of the error processes that affect Optical Mapping data. As such, I begin with a detailed description of these processes and the statistical distributions by which we model them.

Shearing

For all but the smallest of genomes, the Optical Mapping platform does not generate single-molecule restriction maps of the genome in its entirety [21]. Rather, as the DNA molecules are physically manipulated, they break randomly into smaller pieces, each of which results in its own restriction map. We assume that this breakage is

random and that the individual molecule maps' origins are distributed uniformly over the underlying genome (i.e. a Lander-Waterman process [22]). We typically remove very short maps (size less than 200 kb) from consideration; the remaining maps' sizes are consistent with a truncated exponential distribution whose mean is usually around 500 kb.

Missing Cuts

Sometimes a restriction cut does not appear in a single-molecule map even though the restriction enzyme's cognate sequence is present . This could be due to a number of processes, including incomplete digestion, lack of DNA relaxation around a cut, or an error by the machine vision software. We model each restriction cut as a Bernoulli trial with a given probability of success, usually estimated at around $p = 0.8$ [23–27].

Extra Cuts

Occasionally, a restriction cut appears in a single-molecule map at a place where the restriction enzyme's cognate site is *not* present. This could be due to random physical breakage as the DNA is deposited, photo-induced cleavage from laser illumination of the dye with which the DNA is stained, or non-specific cleavage by the restriction enzyme. We assume that these extra cuts occur randomly, and model them Poisson process whose rate parameter is usually $\psi = 0.005$ extra cuts/kb [23–27].

Sizing Error

Restriction fragment sizes are estimated by first computing the integrated fluorescence intensity of the fragment, then scaling the integrated fluorescence intensity by a factor computed from the fluorescence intensity of nearby standards of known size. Because

the staining process is somewhat inconsistent (see Chapter 2), we model the observed size of a fragment with an actual size μ as a draw from a normal distribution $\mathcal{N}(\mu, V(\mu, \mu^2))$ where

$$V(\mu, \mu^2) = \sigma^2(\tau^2 + 1)\mu + \tau^2\mu^2$$

[23, 26–28].

Small Fragment Desorption

Because the electrostatic force binding a DNA molecule to the Optical Mapping surface is a function of the length of the molecule, small restriction fragments (smaller than 5 kb or so) can desorb from the surface and float away; these fragment are then lost from the restriction map. We model the probability that a fragment of size s has desorbed with the exponential distribution $Exp(Log_2(0.5)/\theta * s)$. The distribution is parameterized this way so as to make θ the distribution's median: fragments of size θ desorb with a probability of 50%. The median desorption size is usually estimated at $\theta = 1.35$ kb [20, 26–28].

Hidden Markov Model Implementation

The hidden Markov model I created to analyze Optical Mapping data is presented in Figure 4.3. The model represents the genome being analyzed (i.e. the genome that produced the maps in the Optical Mapping data set); its output symbols are restriction fragment lengths measured in kilobases. It is important to note that, while many HMMs used in bioinformatics have a discrete, finite alphabet of possible output symbols (e.g. the 4 nucleotides of DNA or the 22 peptides in most proteins), this HMM deals with symbols from a continuous domain (e.g. any possible restriction

fragment size.)

The model is built from a reference restriction map, usually but not necessarily generated *in silico* from a reference genome sequence. The model is constructed by first creating a set of *cut states*, $\mathcal{C}$, where each $C_i \in \mathcal{C}$ represents a cut in the reference map. Thus, if the reference map has n cuts, then $\mathcal{C}$ has n elements. The cut states are silent: they produce no output symbols but are instead a representational convenience.

Next, we construct a set of *fragment states*, $\mathcal{F}$. Each state $F_{i,j}[k] \in \mathcal{F}$ represents the possibility that the cuts between C_i and C_j are missing, but instead there are $k-1$ random extra cuts (i.e. k output fragments). Concrete examples are presented in Figure 4.4. The transition probability from C_i to $F_{i,j}[k]$ is simply the product of the probability that the cuts between C_i and C_j are missing, and the probability that $k-1$ extra cuts were observed in the region between C_i and C_j: these values are readily calculated from the probability distributions of the error models presented above. (Because there is only one transition out of $F_{i,j}[k]$, to C_j, its probability is 1.0, and are represented in the implementation only implicitly.) Such a construction also provides a convenient way to reign in the size of the model which would otherwise grow exponentially with the size of the reference: we can instead set a probability threshold, say 10^{-4}, and then only build fragment states whose incoming transition probability is above that threshold.

Thus are the HMM transition probabilities used to model the extra and missing cut errors. Sizing errors, on the other hand, are modeled by the emission distributions associated with each fragment state. The fragments produced by state $F_{i,j}[k]$ are the result of drawing a random observation from the sizing error model $\mathcal{N}(\mu, V(\mu, \mu^2))$ presented above, then randomly breaking that draw into k pieces.

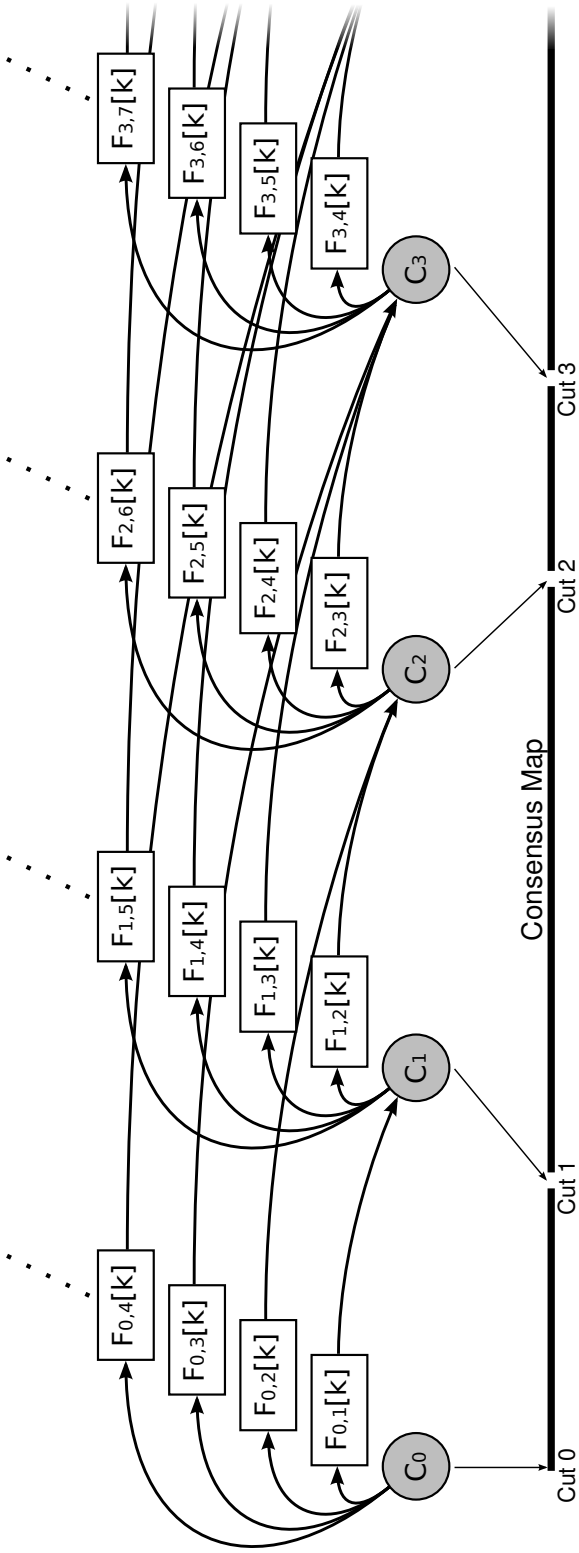

Figure 4.3. A hidden Markov model for Optical Mapping data analysis. Each fragment state $F_{i,j}[1], F_{i,j}[2], ...$ has the same connectivity; for clarity, they are collapsed into a state labelled $F_{i,j}[k]$.

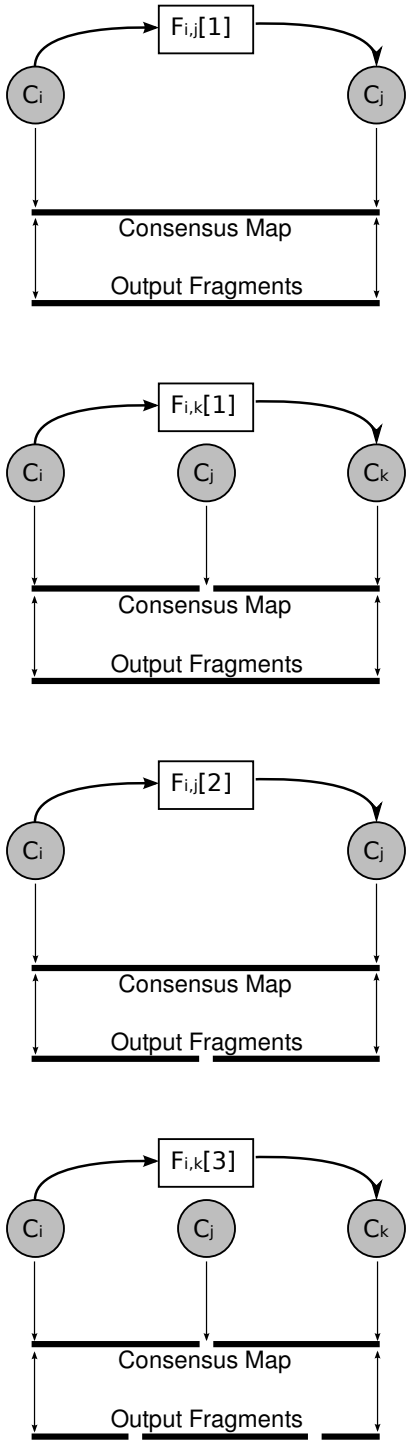

Figure 4.4. Different HMM states represent different Optical Mapping errors.

The last remaining error process, small fragment desorption, is modeled by the addition of fragment states that don't produce any fragments. State $F_{i,j}[0]$ represents the possibility that all the restriction cuts between C_i and C_j are missing, and that the resulting fragment desorbed. The incoming transition probability is the product of the probability that the cuts between C_i and C_j are missing, and the probability that a fragment of that size desorbed; and like before, only states for which this probability is above some minimum threshold are created when constructing the model.

Thus, a single-molecule restriction map generated by the genome from which we built our model can be represented by a path through the model (Figure 4.5). In fact, it can usually be represented by many different paths: a restriction map with n fragments could have been the result of any sequence of states whose number of output fragments sum to n (though with different probabilities of being produced by the different paths). Thus, two of the natural questions are "Which sequence of states was most likely to have produced the map?" and "What is the sum of the probabilities of all the possible paths that might have produced the map? In other words, what is the probability that the restriction map was produced by the model?" I return to these questions in the following sections.

Gaps

Though not a direct result of any of the Optical Mapping error processes described above, it is frequently useful from an analytical standpoint for our model to consider gaps as well. Gaps can represent larger changes between the reference from which we built the model and the genome from which we generated the single molecule restriction maps we're analyzing. As with traditional gapped sequence alignment,

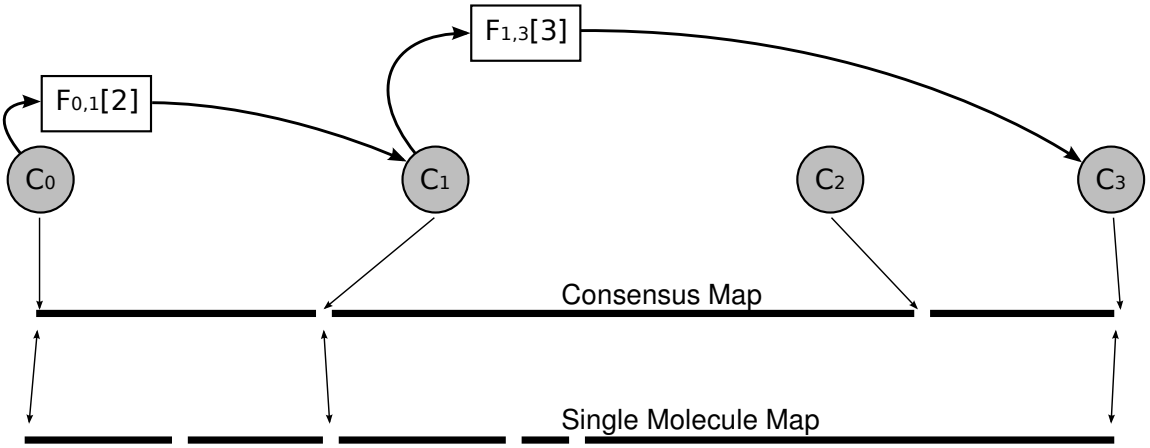

Figure 4.5. A single-molecule restriction map is represented by the path through the model that generated it.

the gap penalty is separated into the penalty for opening a gap and the penalty for extending it. Here, the gap extension penalties are applied per-fragment (instead of, say, per-kilobase) and are different between missing reference map fragments and missing single-molecule restriction map fragments.

We model gaps with a new set of output states $\mathcal{G}$, where $G_{i,j} \in \mathcal{G}$ represents a gap between model cuts C_i and C_j, where C_i and C_j might be the same cut. The transition probability from C_i to $G_{i,j}$ is the product of the probability of opening a gap and the penalty for the fragments between C_i and C_j being missing. The state $G_{i,j}$ can emit a variable number of fragments, from 0 to some prespecified maximum; the emission probability for emitting k fragments is the penalty for that number of missing fragments in the single-molecule restriction map.

4.2 Common HMM Algorithms Validate the Optical Mapping HMM

The hidden Markov model described above provides a well-grounded probabilistic framework with which to analyze Optical Mapping data. Actually performing these analyses, however, requires additional algorithms that use the data structure to perform inference. Happily, a set of efficient algorithms exist with which to solve the three most common inference problems: evaluation, decoding and learning. These three inference problems are important because they correspond to common analyses we wish to perform on Optical Mapping data. They also serve to validate the model before we embark on less well-studied problems: testing these algorithms with synthetic data allows us to evaluate the model's performance when the ground truth is known, providing reassurance that the model is performing properly.

Reference Discrimination (Evaluation)

As discussed previously, one common use for hidden Markov models is in the evaluation of input sequences: "What is the probability that the model produced a given sequence?" This question is especially useful when comparing multiple models in a model-selection context: then the question becomes "Which of these different models most likely produced the given sequence of observations?" This inference is commonly implemented with the Forward algorithm [6, 29], a dynamic programming algorithm that runs in $O(mn)$ time, where m is the size of the model and n is the size of the input.

To test the model's performance, I selected two arbitrary 5-megabase regions

of the human genome reference map and generated 500 synthetic single-molecule maps from each. Then, I constructed an HMM for both references and computed the probability that HMM had produced each synthetic single-molecule map. I converted these probabilities into log-odds scores, dividing each by the mean probability of 10 random permutations of that map. The results are plotted in Figure 4.6 as a function of map size (in fragments): the blue crosses represent the probability computed by the HMM that generated the map ("match") while the red crosses represent the probability computed by the HMM that didn't ("mismatch"). The HMMs discriminate between matched and mismatched synthetic maps almost perfectly. Also of note, the discriminative power increases as the synthetic map size increases. (No surprise, as longer maps have a greater information content with which to perform inference.)

Pairwise Alignment (Decoding)

Another common use for hidden Markov models is in decoding input sequences: "What sequence of model states is most likely to have produced a given sequence of observations?" This inference is commonly implemented with the Viterbi algorithm [6, 29, 30], a dynamic programming algorithm that runs in $O(mn)$ time, where m is the size of the model and n is the size of the input. Importantly, solving the decoding problem corresponds directly to the pairwise alignment between the reference map and a single-molecule restriction map, which is as common in analyzing Optical Mapping data as it is in analyzing nucleotide or polypeptide sequences.

To test the model's performance, I generated 500 synthetic single-molecule maps from a reference map of the *Y. pestis* genome. Because data was generated synthetically, the true location which produced each synthetic map is known. Then,

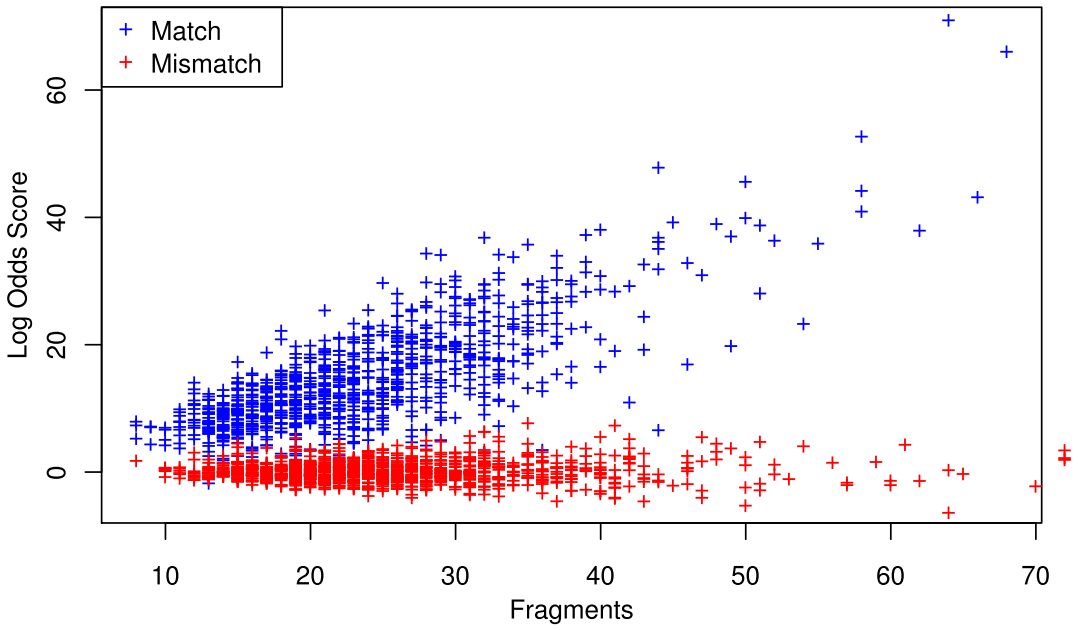

Figure 4.6. The HMM discriminates between correct and incorrect models. The plot represents the log-odds scores of two synthetic mapsets, evaluated by the HMM for the reference that produced them ("Match") and by the HMM for the reference that didn't ("Mismatch"). The HMMs discriminate between matched and mismatched synthetic maps almost perfectly.

I constructed an HMM from the *Y. pestis* reference map and used it to perform pairwise alignments of the synthetic maps, asking whether the HMM could decode where in the reference the synthetic maps had come from. The results are plotted in Figure 4.7; alignments covering at least 70% of the original location are plotted in blue, while incorrect alignments are plotted in red. Of the 500 synthetic maps, 485 (97%) are correctly aligned by the HMM. Incorrectly aligned maps tended to be smaller, explaining the difficulty the HMM had in properly decoding them.

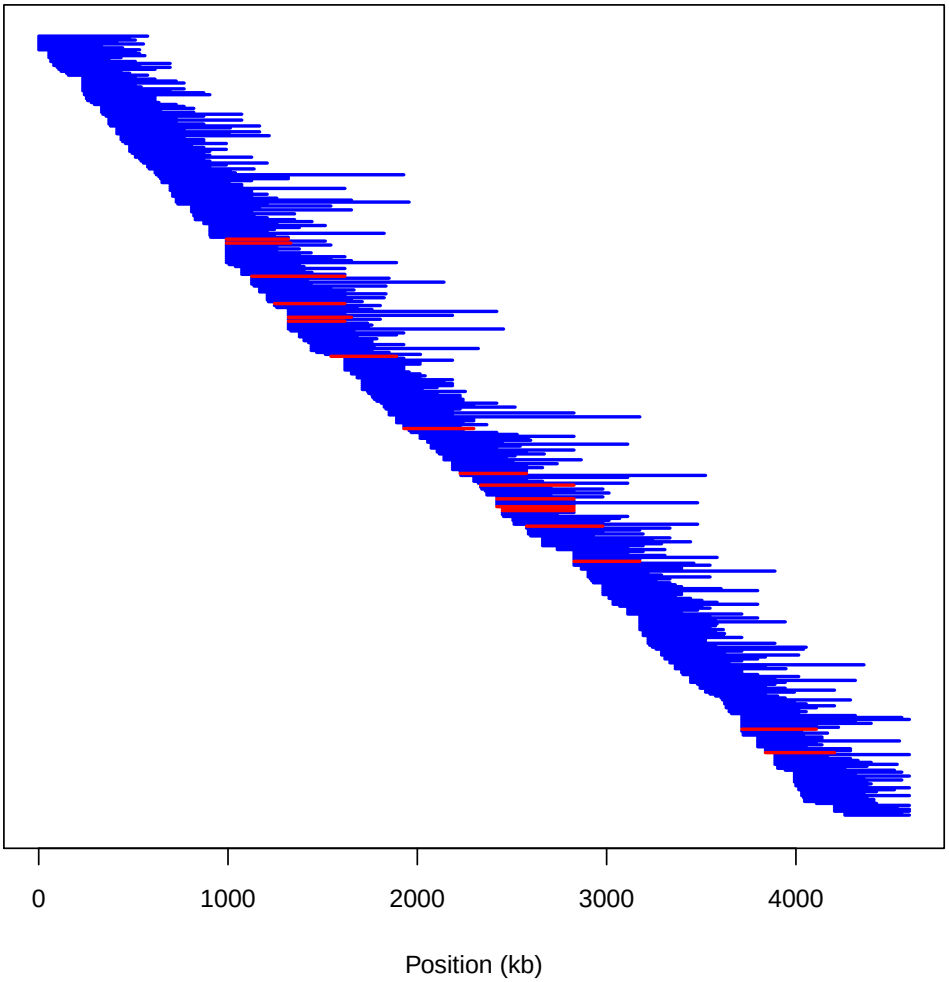

Figure 4.7. The HMM can be used for pairwise alignment. The ideogram represents the pairwise alignments between 500 synthetic single-molecule maps and a reference. Each horizontal line represents a single-molecule map; blue lines are maps that aligned to the correct location, while red lines are incorrectly aligned maps. The hidden Markov model was highly successful in aligning synthetic maps back to the correct location.

Parameter Estimation (Learning)

The previous two subsections discussed inference performed on the *data*; we turn now to inference performed on the *model.* In particular, we are interested in learning the parameters (transition and emission distributions) of the model, answering the question "Which parameters make the model most likely to have produced the data we actually observed?"

This inference is commonly implemented on HMMs with the Baum-Welch algorithm [6, 29, 31]. In this instance, though, some modification must be made because the transition and emission distributions are actually parametric (i.e. the distributions discussed in the error models.) So, rather than infering the best transition and emission probabilities, what we want to infer are the best error model parameters: restriction digest efficiency, false cut rate, sizing error parameters and median missing fragment size. Below, we focus on digestion efficiency and false cut rate, though sizing error parameters and median missing fragment size are amenable to similar approaches.

Like the pairwise alignment problem, our approach is to use posterior decoding, but phrased in a different way: instead of "What is the sequence of states most likely to have produced this sequence of observations," we ask "What is the probability that state S produced one of the observations in the current sequence?" This posterior probability, summed for state S across all the observations in a data set, approximates the number of times that state S "fired," or was involved in producing an observed piece of data. Thus, for a given cut state, we can compute the proportion of observations that went through the cut state, versus through a fragment state in which that cut was missing. This proportion is the efficiency with which that

restriction cut was actually cleaved. Averaging across the entire reference map results in an estimate of the cutting efficiency for the entire data set. A similar procedure can be used to estimate the false cut rate as well. The estimation procedure is a fixed-point algorithm, iterating cycles of posterior decoding and parameter estimation until the estimate converges.

As above, we use synthetic data to test the performance of the model at this inferential task. In this instance, I generated multiple synthetic data sets from the *Y. pestis* genome with a known digestion rate and a known false cut rate. Then, I ran the estimation procedure for digestion rate and false cut rate several times, starting each time with a different value for the parameter. As Figure 4.8 shows, no matter where the relevant parameter starts, its estimate converges to the actual value.

4.3 Reference Map Refinement *via* HMM Model Selection

As described in Chapter 3, our laboratory uses an interative procedure [32] to assemble consensus restriction maps spanning the entire genome. Frequently, the goal is to compare these consensus maps back to a reference genome to study how the genome represented by the consensus maps differs from the reference. Unfortunately, the iterative assembly procedure is quite slow for mammalian-sized genomes. Gentig [23, 33], one of its key components, is a Bayesian *de novo* assembly engine that runs in time that on average scales as $O(n \times MlogM)$, where M is the number of input maps and n is the average number of fragments per map. While the availability of grid-computing resources [34, 35] ameliorates this concern somewhat, future Optical

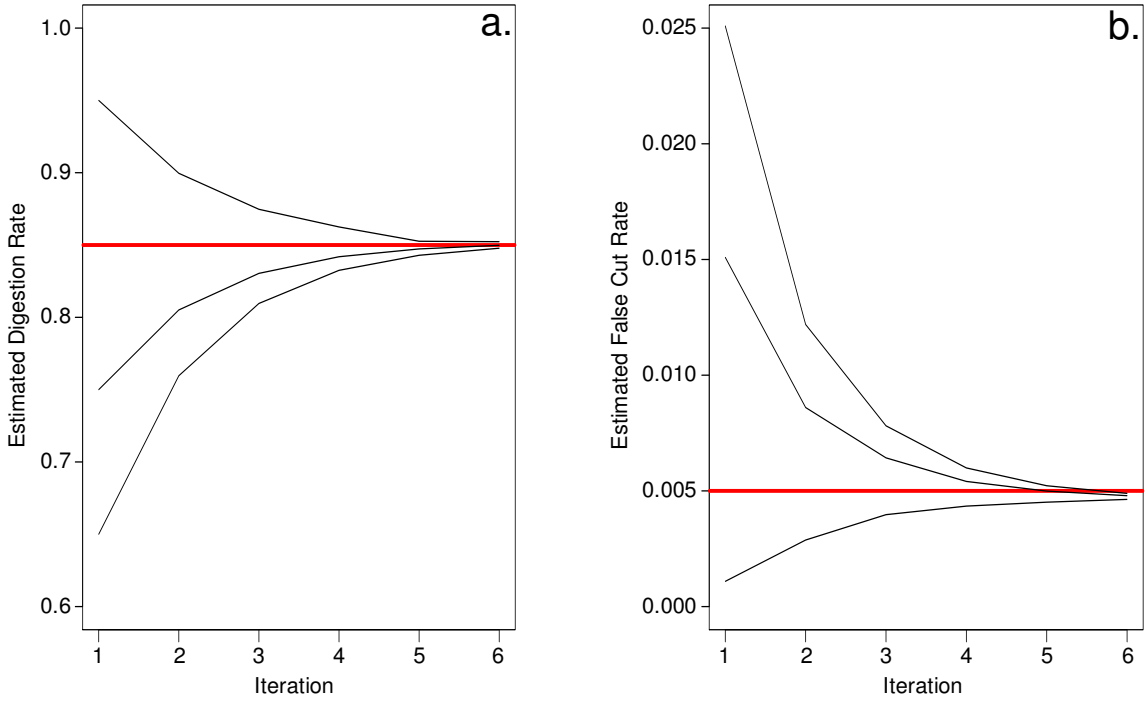

Figure 4.8. The hidden Markov model can estimate unknown error model parameters. A synthetic data set was created with a known enzyme digestion rate (panel a) or a known false cut rate (panel b), represented in each case by the red horizontal line.. The iterative estimation procedure was started with several different initial parameterizations; in each case, the estimate converged to the correct value.

Mapping projects are likely to be constrained by computational concerns rather than experimental ones.

Fortunately, the ease with which a hidden Markov model can solve the "evaluation" problem suggests an alternate approach. Rather than doing an entire assembly *de novo* and then comparing it to the reference, perhaps it would be possible to find those differences by simply asking what model is most likely, given the data we observed? If we want to maximize $P(model|data)$, then Bayes' Law lets us do so by instead maximizing $P(data|model) \times P(model)$. Because a fast algorithm exists to compute $P(data|model)$ on an HMM, the problem becomes one of searching model space for

models that increase the probability of observing the data. The following sections describe this approach in detail and analyze its performance, both on synthetic data for which the ground truth is known and on a real Optical Mapping data set as compared to iterative assembly.

The idea of searching model space to arrive at a consensus map that best represents an Optical Mapping data set is not a new one; similar (though discretized) approachs have been used derive consensus restriction maps of clones [24, 36], and a search through a high-dimensional model space is at the core of Gentig [23, 33]. The key insight here is to use a pre-built model based on a pre-existing reference sequence to constrain and guide the model search, delivering a significant speedup.

Approach

As introduced above, the goal is to alter the hidden Markov model constructed from a reference map so as to maximize $P(model|maps)$. By Bayes' Law, we can convert this to a maximization of $P(maps|model) \times P(model)$. Thus, as $P(maps|model) \times P(model)$ increases, the hidden Markov model and the reference restriction map which it represents become a better representation of the genome that produced the single-molecule restriction maps. But what possible changes should we test?

The posterior decoding described for parameter estimation provides a convenient way to choose candidate changes. For each fragment state $F_{i,j}[k]$, we can compute the sum of the posterior probabilities of all the single-molecule restriction fragment in the data set. Put less rigorously, we compute how many fragments in the data set were "produced" by each fragment state. Fragment states with a sum-of-posteriors greater than some threshold become candidate changes to test to see if they increase

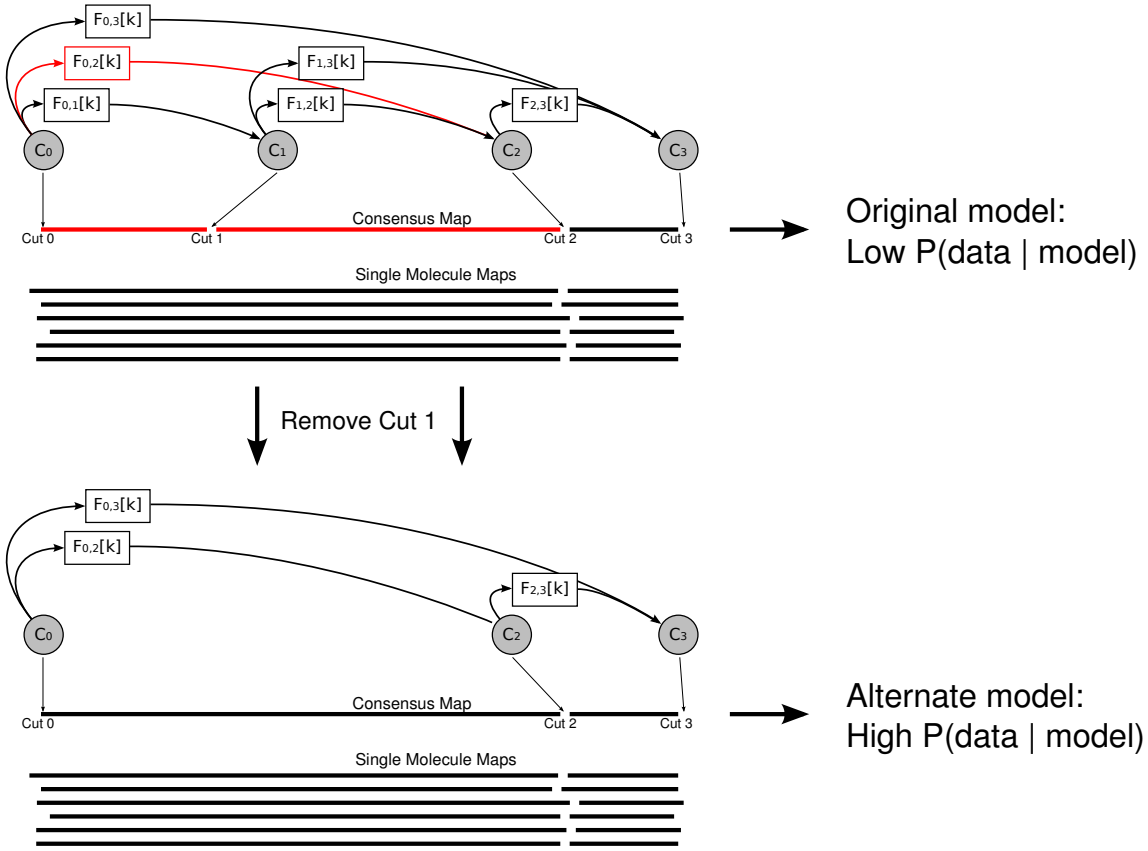

Figure 4.9. Using a hidden Markov model to refine a consensus restriction map.

$P(maps|model)$. For example, fragment state $F_{0,2}[1]$ connects cuts C_0 and C_2 with only one fragment; in other words, a map produced by a path through $F_{0,2}[1]$ has reference cut C_1 missing. If we replace the part of the reference between cuts C_0 and C_2 with the one fragment produced by fragment state $F_{0,2}[1]$, we can rapidly recompute $P(maps|alternate model)$ to measure what effect removing cut C_1 from the reference has on the overall probability that the model produced the single-molecule maps that were actually observed (Figure 4.9).

When making this replacement, what fragment size should we use to go between C_1 and C_3? Posterior decoding can be of use here, too: we can keep a list of fragments

produced by $F_{1,3}[1]$, then weight them by their respective posteriors and use those weighted observations to find the fragment size whose sizing error model distribution $\mathcal{N}(\mu, V(\mu, \mu^2))$ was most likely to have produced those observations. Because the distribution's variance takes a complicated form, I elected to use quasi-Newton minimizer [37] to find the best μ instead of attempting to do so analytically.

Once the probabilites $P(maps|alternate\,model)$ have been computed for all the candidate changes to the model, it remains to select the set of changes that result in the largest cumulative increase in $P(maps|model)$. I begin by storing the effect of any change $Change_i$ as $R_i = P(Change_i) \times P(maps|model\ with\ Change_i)\ /\ P(maps|original\ model)$ where $P(Change_i)$ is the prior probability of introducing $Change_i$. Then, I make a key simplifying assumption:

$$\frac{P(maps|model\ with\ Change_i\ and\ Change_j)}{P(maps|original\ model)} = \frac{P(maps|model\ with\ Change_i)}{P(maps|original\ model)} \times \frac{P(maps|model\ with\ Change_j)}{P(maps|original\ model)}$$

That is, if making $Change_i$ doubles $P(maps|model)$, and making $Change_j$ *also* doubles $P(maps|model)$, then making *both* $Change_i$ and $Change_j$ will *quadruple* $P(maps|model)$. In other words, the effects of $Change_i$ and $Change_j$ on $P(maps|model)$ are independent of one another. Under this assumption, one simply chooses the sequence of adjacent changes $Change_i$, $Change_j$, $Change_k$... so as to maximize the product $R_i \times R_j \times R_k$.... (By converting all R_i to negative log space, Dijkstra's shortest paths algorithm can be used.) In practice, assuming that $Change_i$ and $Change_j$ affect $P(maps|model)$ independently sometimes results in spurious changes to the model; better model updates can be had by considering adjacent *pairs* of fragment states as a single change, at the cost increasing the number of changes that must be considered.

Model Selection Priors

An important consideration in any Bayesian approach is selection of a prior. How do we choose the prior probability $P(Change_i)$ for the change induced by some fragment output state $F_{i,j}[k]$? Naïvely assuming a uniform prior results in widespread overfitting, even when no change exists. For example, given data that precisely matches the model between cuts C_1 and C_3, the algorithm above should choose a sequence of "changes" induced by $F_{1,2}[1]$ and $F_{2,3}[1]$ (i.e. a new model that represents no change from the old one.) Instead, it frequently chooses the change induced by $F_{1,3}[2]$ because the minimization procedure that chooses the fragment sizes for $F_{1,3}[2]$ can produce fragments that "outperform" those estimated for $F_{1,2}[1]$ and $F_{2,3}[1]$.

Instead, we want a prior that matches our intuition that changing the model is less likely than keeping it the same, and that larger changes are less likely than small changes. This intuition is precisely matched by the transition probabilities from C_i to $F_{i,j}[k]$, as described previously. In practice, even this prior does not sufficiently control over-fitting. Instead, we take a page from the Bayesian model selection literature [38–40] and take into consideration the amount of data that goes into the model choice: if the transition probability from C_i to $F_{i,j}[k]$ is t, then the prior for $Change_i$ induced by $F_{i,j}[k]$ is

$$t^{1+d\times s}$$

where d is the depth of the single-molecule maps at that locus, and s is a small multiplier usually set at 0.1 or 0.2. Thus, s provides a parameter that controls the willingness of the model to call a variant, at the cost of increasing false-positives from over-fitting.

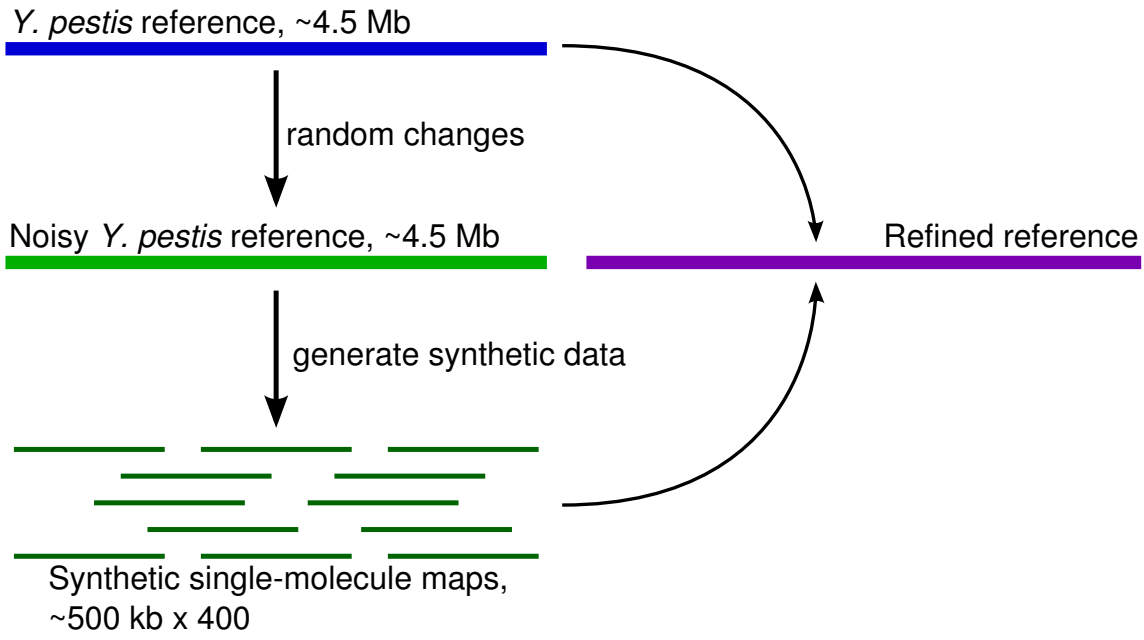

Figure 4.10. Using synthetic data to validate the map refinement algorithm.

Validation with Synthetic Data

Before applying this approach to real data, I tested it on a synthetic data set. Because the ground truth is known *a priori*, this allowed for a rigorous evaluation of the approach's performance. The experimental approach is outlined in Figure 4.10. I began with a 4.6 Mb reference genome restriction map generated *in silico* from the *Y. pestis* reference sequence, then introduced random extra cuts, missing cuts and sizing errors to create a "noisy" reference. I used this noisy reference to generate a synthetic Optical Mapping data set consisting of 400 synthetic restriction maps, of average size 500 kb, for an average 40-fold coverage of the reference genome. I then applied the refinement procedure, starting with the original reference and the synthetic data set, to see if the algorithm could recover the changes that had been introduced into the "noisy" references.

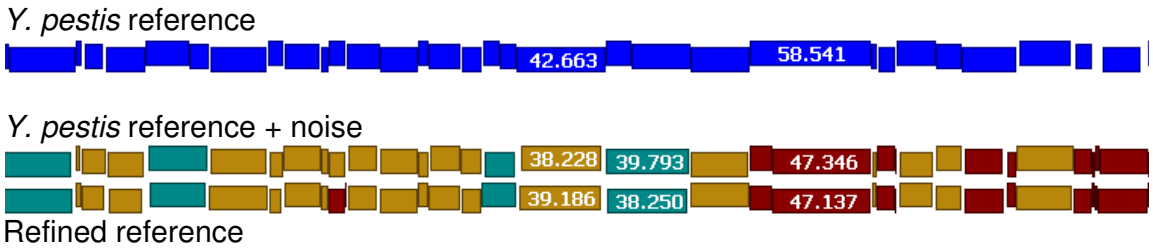

Figure 4.11. The HMM-based refinement procedure correctly recovers changes to the reference. The alignments show an example region from the *Y. pestis* genome, showing that the HMM-based refinement procedure recovers most of the differences introduced by randomly perturbing the reference map.

An example region is presented in Figure 4.11. The refined model almost perfectly recapitulates the noisy reference: of the 31 new extra cuts, the refinement procedure recovered 30 with 2 false positives. Of 23 new missing cuts, the refinement procedure recovered all 23 with no false positives. Additionally, 28 fragments were resized to be at least 10% smaller or larger; all 28 are recovered as significant insertions or deletions (Figure 4.12).

Comparison to Iterative Assembly

A laboratory colleague, Aditya Gupta, has collected a high-quality Optical Mapping data set from a matched trio of multiple myeloma cancer samples. Multiple myeloma is characterized by an overproliferation of plasma cells in the blood; the primary malignancy is generally responsive to chemotherapy [41] (frequently paired with an autologous stem-cell transplant [42]), but invariably relapses and becomes treatment-resistant. The trio of tissue samples from which Aditya created Optical Mapping data sets are a cell line derived from normal tissue; plasma cells collected before relapse and treatment-resistance; and post-resistance plasma cells. The biological relevance of these samples is complemented by exceptionally high-quality data, providing an

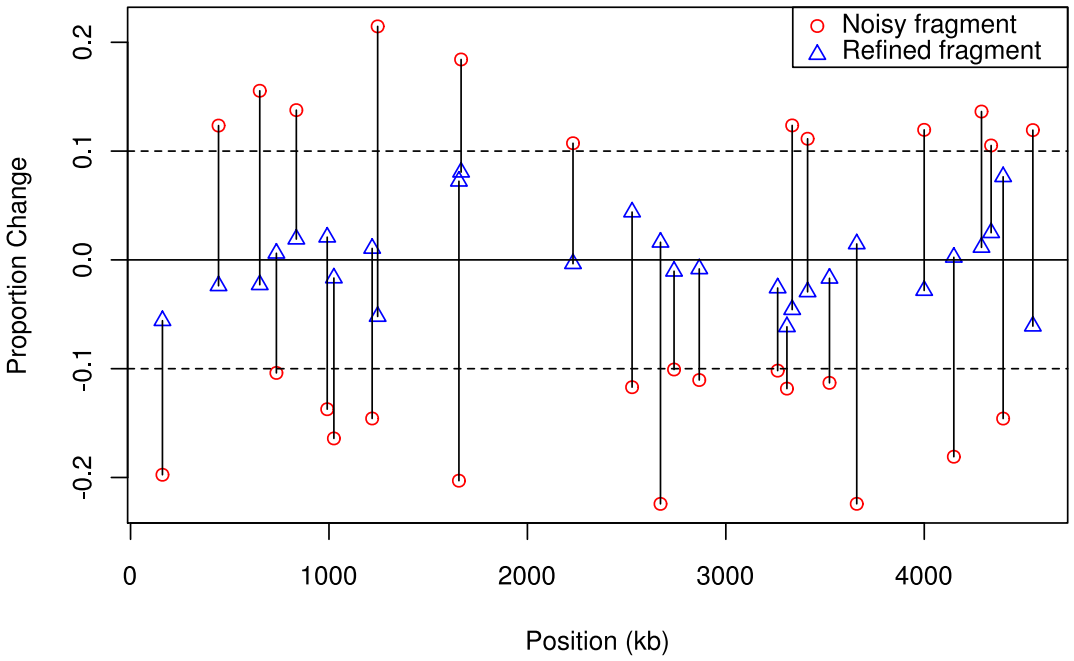

Figure 4.12. The HMM-based refinement procedure correctly estimates changed fragment sizes. Twenty eight fragments were changed to be over 10% larger or smaller than their original size, simulating insertions or deletions in the noisy reference relative to the original. All 28 of these differences were recovered by the refinement procedure.

unparalleled substrate on which to test this new algorithm's performance.

I performed the hidden Markov model-based map refinement on chromosome 13 of the normal tissue, of biological interest because whole-chromosome deletions of chromosome 13 have been associated with poor prognosis [43]. Then, I compared my results to a previous iterative assembly analysis (Figure 4.13 and Table 4.1). Between the two analyses, 69 differences were detected between the experimental genome and the NCBI Build 37 reference sequence [44]. Forty two of those differences were found

	Found by HMM	Not found by HMM
Found by Iterative Assembly	42	20
Not found by Iterative Assembly	7	

Table 4.1. Performance of iterative assembly and HMM-based refinement compared. A majority of the variants from multiple myeloma chromosome 13 were found by both methods. Of the variants that were not shared, a majority appear heterozygous.

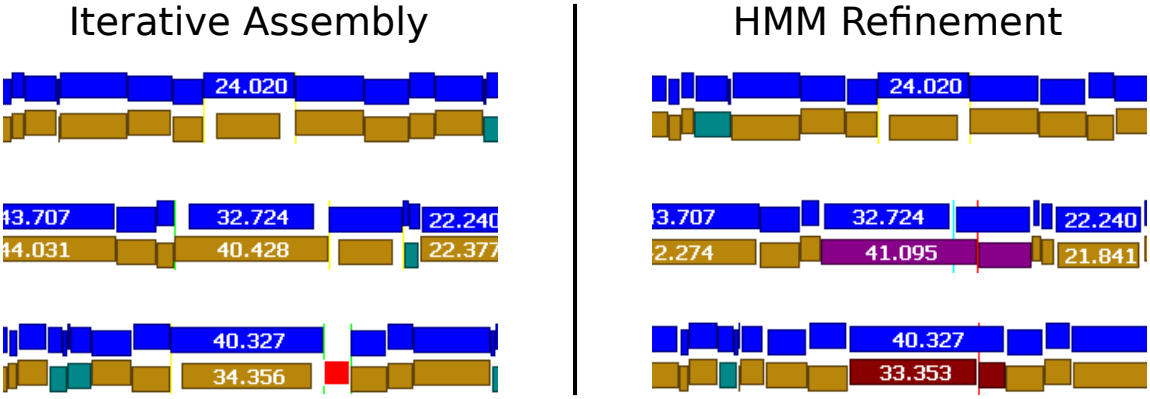

Figure 4.13. For many variants, iterative assembly and HMM-based refinement perform similarly. The three example loci show that, even if the precise nature of the variant wasn't the same, the underlying structural change is usually found by both methods.

by both methods; 7 were found only by the HMM refinement, and 20 were found only by iterative assembly. Of note, 4 of the 7 HMM-only variants appear to be heterozygous loci, while 18 of the 20 iterative assembly-only variants appear to be heterozygous (Figure 4.14). Also of note, the runtime of the HMM-based refinement method was substantially better, requiring 114 CPU-days of computation versus the 347 CPU-days required by the iterative assembly method.

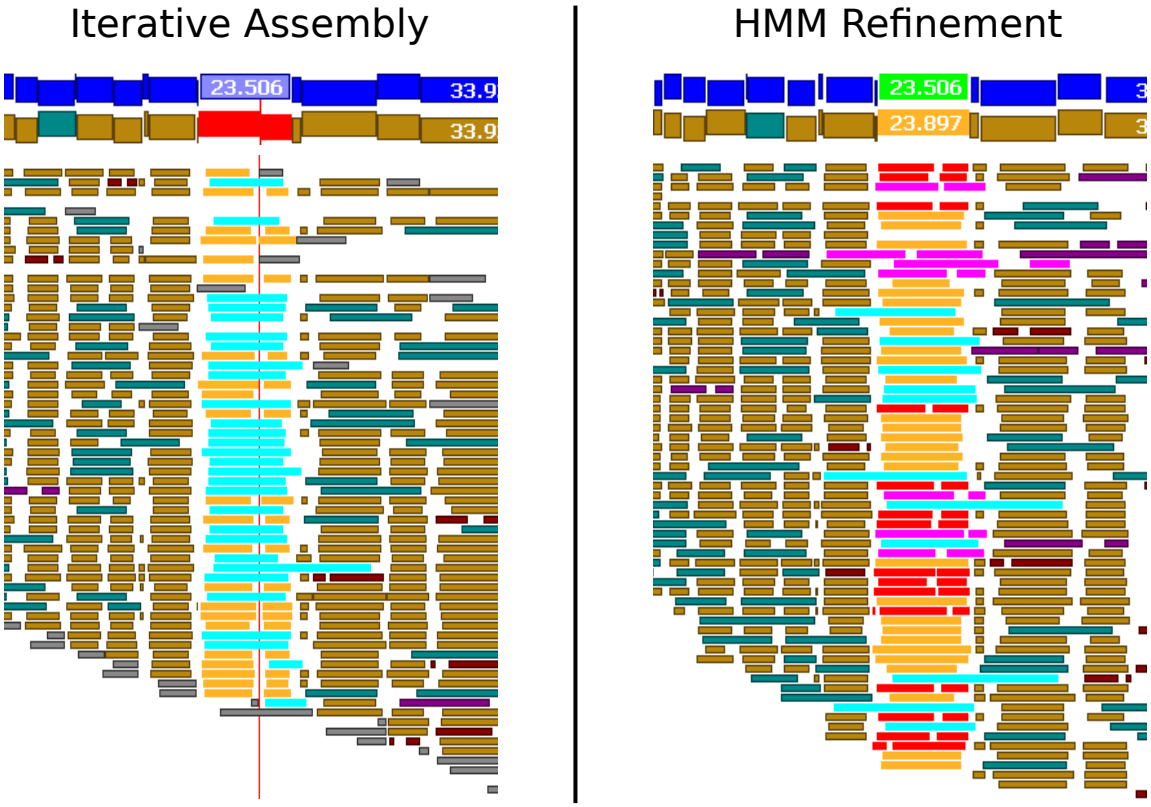

Figure 4.14. Many of the variants found by one method but not the other appear to be heterozygous. The consensus-to-reference alignments are presented above, and portion of the single-molecule restriction maps aligned to the locus are presented below. There appears to be a bimodal distribution of fragments in the highlighted column; the iterative assembly procedure calls an extra cut, while the HMM-based refinement procedure does not.

4.4 Haplotype Recovery with Belief Propagation

The number of potentially heterozygous variants in the previous results highlight another problem with current analysis methods: because our assembly engine is designed to operate on clonal samples [23, 33], our current discernment of heterozygosity is serendipitous, or at best ad-hoc. The accurate disambiguation of haplotypes in a diploid sample would be a major advancement in our ability to derive biological

meaning from Optical Mapping data sets.

Unfortunately, it is far from clear how best to separate maps from mixed samples in an Optical Mapping data set. Two approaches suggest themselves. The first is locus-by-locus: at each point in an optical map assembly, one could perform some test to distinguish whether the single-molecule maps at that position appear to have come from one underlying genome or two. This approach is intuitive: when scanning the genome assembly visually, this is how one recognizes a possibly heterozygous site. Also, because it's local, performing this test can be made fast, even though it has to be repeated at every locus in the genome.

However, there are problems with a local approach. First, just how does one define "locus?" If one is satisfied with reporting heterozygosity in simple missing- or extra-cuts, then perhaps a single cut site or restriction fragment is sufficient. However, if one is interested in larger-scale differences, or in recovering phasing between two linked loci, the locus-by-locus approach rapidly gives way to a global assignment problem: if one could assign each map to one contributing genome or the other, then previous methods could be used to recover per-haplotype variants. This strategy also recovers phasing: two variants supported by the same underlying single-molecule maps should be assigned to the same haplotype.

The major drawback to a global assignment problem is the computational effort it requires to solve: as currently formulated, the assignment problem is an optimization problem with one variable for each map and one variable for each candidate fragment output state. Gibbs sampling [45] offers one possible solution for determing the marginal distributions of all of these variables, but such approaches converge slowly and are not gracefully parallelized. A natural alternative is belief propagation on a factor graph; this is the approach presented below.

Belief Propagation in a Nutshell

Belief propagation is a message passing algorithm for performing inference on graphical models [46–49]. It allows for the efficient computation of marginal distributions of variables on a factor graph (Figure 4.15): for example, given some joint probability function $P(x, y, z)$ that can be *factored* into a number of (more readily computed) functions $P(x, y, z) = \phi_1(x, y) \times \phi_2(x, y, z) \times \phi_3(y, z) \times \phi_4(z)$, we construct a variable node for each variable $x,\ y,\ z$ and a factor node for each factor $\phi_1,\ \phi_2,\ \phi_3,\ \phi_4$. We add edges between variables and the factors in which they appear. The algorithm proceeds by passing messages along edges between the nodes, each containing the "influence" a factor exerts on a variable and *vice versa*. If the factor graph is a tree (i.e. has no cycles), then the algorithm takes only one message passed in each direction on each edge to solve for the extact marginal distributions of the variables [46]. However, the algorithm has been shown to be a useful approximation even on general, loop-containing graphs [49, 50].

Approach

In the map refinement section previously, the goal was to optimize $P(model|data)$. Applying Bayes' Law, I used a relatively simple search through model space to instead optimize $P(data|model) \times P(model)$. For solving the global haplotype assignment problem, a different approach is required: we want to assign haplotypes such that, given two separate HMMs, we maximize $P(data|model) \times P(model)$ for both HMMs. The problem is, the best model given the observed data changes as the haplotype assignments change. If we were systematically exploring haplotype assignment space (say, with a Gibbs sampler), then it would make sense to define our objective function

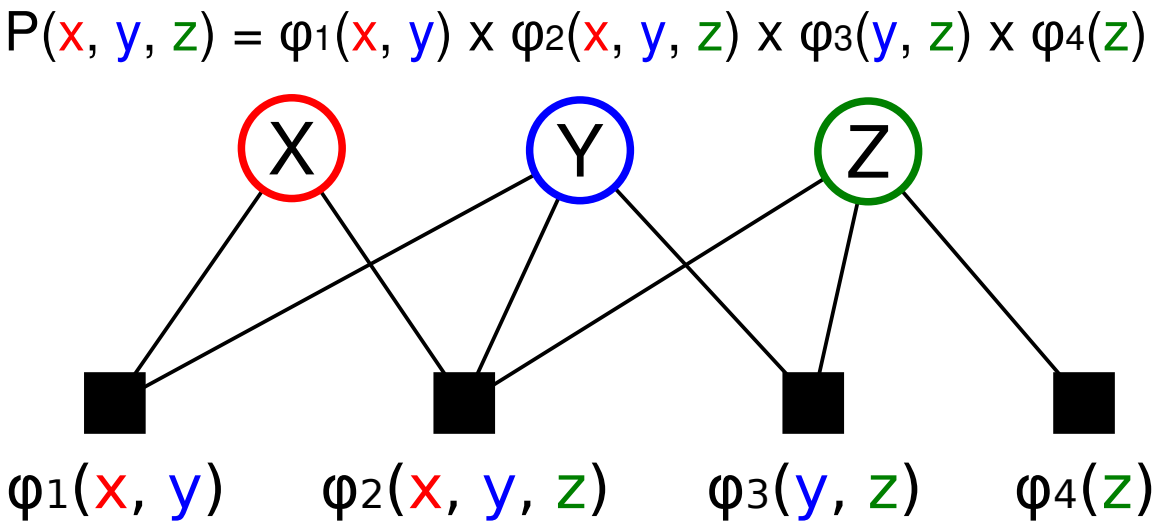

Figure 4.15. An example factor graph for belief propagation. The joint probability function $P(x, y, z)$ is factored into $P(x, y, z) = \phi_1(x, y) \times \phi_2(x, y, z) \times \phi_3(y, z) \times \phi_4(z)$. The factor graph thus induced allows for the efficient computation of the marginal distributions of X, Y and Z.

(not a true joint probability function, because it does not integrate to unity) as the product of the best model's $P(model|data)$ for each haplotype under consideration. Unfortunately, a belief propagation approach works best with a continuous objective function, whereas $P(model|data)$ changes discontinuously as the number of fragments in the model changes.

To overcome this problem, I defined my objective function as the *sum* of $P(model|data)$ *for all possible changes.* That is, given a particular haplotype assignment, I compute the probability of $P(data|model) \times P(model)$ for *all* sequences of candidate changes to the model, selected by the same method as with model refinement, and sum them together. (Because this computation is also structured as a trellis, a dynamic programming algorithm similar to the Forward algorithm on an HMM allows this to be computed rapidly). The objective function is the product of this value for all the haplotype models under consideration.

The benefits of this objective function are twofold. First, as mentioned previously, it is completely continuous. It doesn't matter whether $Change_i$ or $Change_j$ ends up in the "best" new model, because they're being summed together. Second, this objective function penalizes mixtures of haplotypes because any particular change that increases the probability of the data from one haplotype, decreases the probability of the data from the other.

Belief propagation works by passing messages on a graph of factors and variables: in this case, the factor graph consists of one variable vertex and one factor vertex per single-molecule map. Each map's variable vertex represent the distribution of the map's haplotype assignment. To create the factors, I use the chain rule to factor out the contribution of each single-molecule map's haplotype assignment to the overall objective function: if map_i's haplotype assignment to a particular HMM is represented by A_i, and each HMM's portion of the objective function is $P(A_1, ..., A_n)$, then

$$P(A_i|A_1, ..., A_{i-1}, A_{i+1}, ..., A_n) = \frac{P(A_1, ..., A_{i-1}, A_i, A_{i+1}, ..., A_n)}{P(A_1, ..., A_{i-1}, A_{i+1}, ..., A_n)}$$

Repeated for each single-molecule map, this completely factors $P(A_i, ..., A_n)$.

Belief propagation on the above factor graph was implemented as per [48]. If the HMMs modeling the haplotypes are identical (as in the first round of map refinement), then single-molecule map haplotype assignment distributions are chosen randomly to break the symmetry. If the haplotype models are not identical, then the haplotype assignment distributions begin uniformly distributed between haplotypes. Generally, five iterations of message passing seem sufficient to maximize the objective function (Figure 4.16). After maximization, each single-molecule map is assigned to the HMM corresponding with the largest fraction of its haplotype assignment distribution.

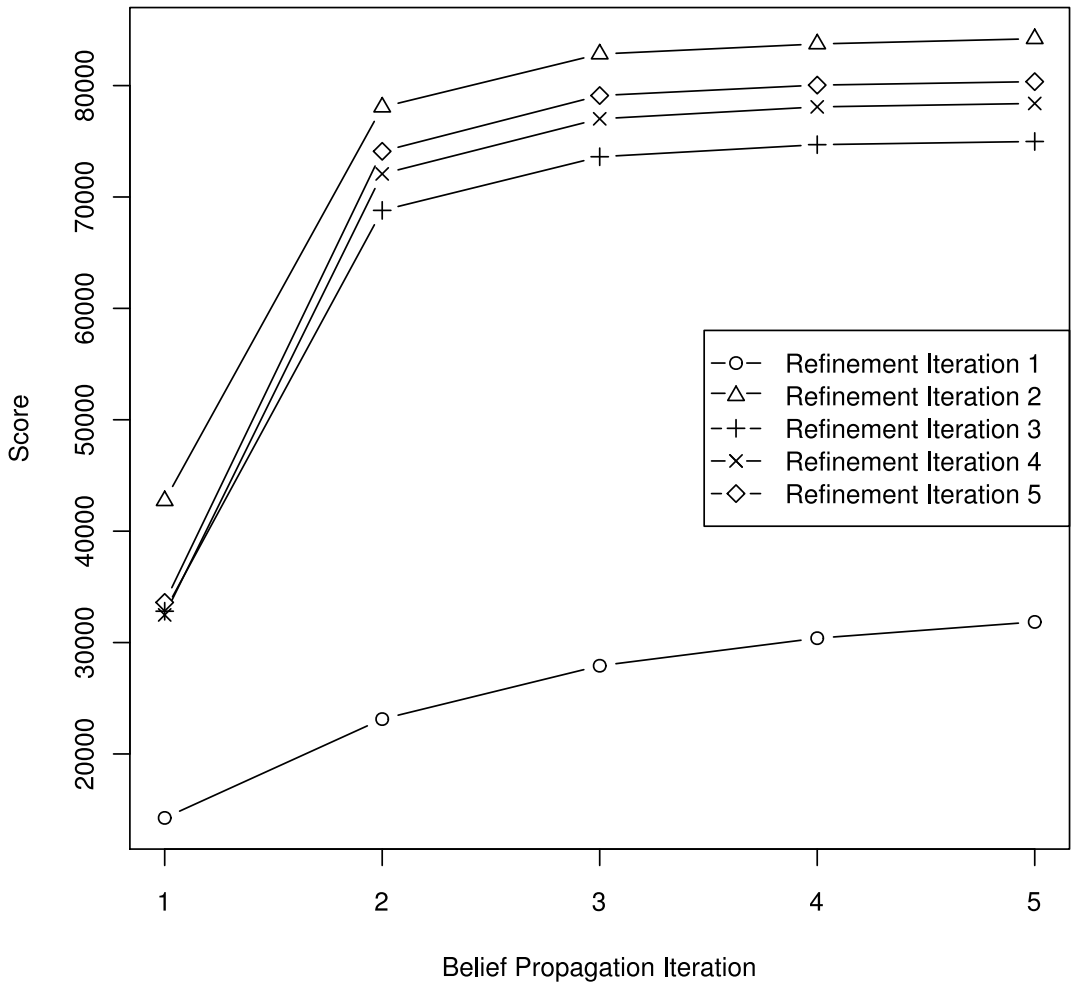

Figure 4.16. Belief propagation maximizes the factor graph's objective function. These data from the multiple myeloma chromosome 13 experiment described below.

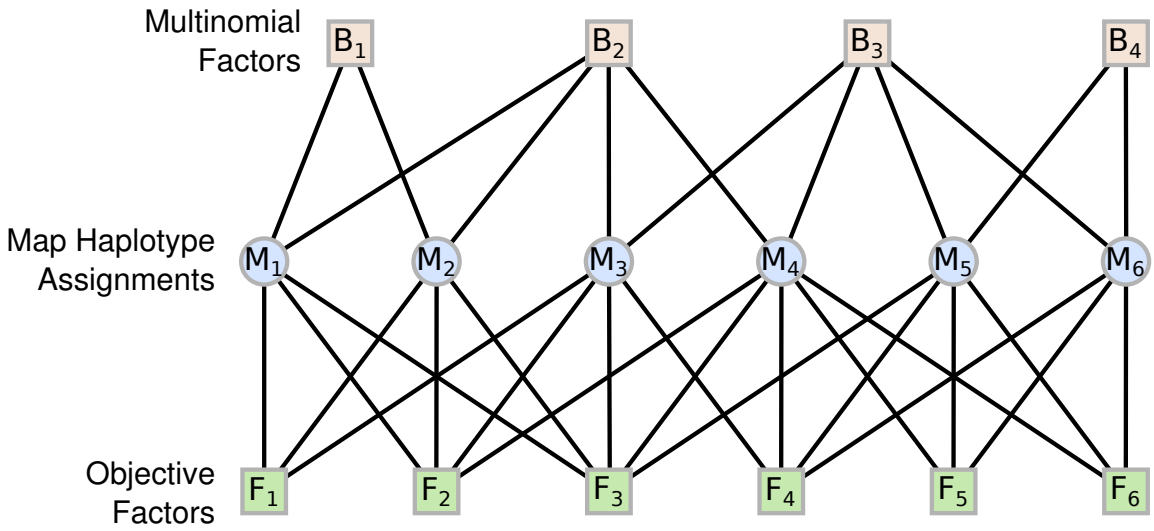

Figure 4.17. The factor graph for using belief propagation to assign single-molecule haplotypes to Optical Mapping data.

The algorithm and factor graph described above usually produce correct output, but occasionally all of the single-molecule maps end up assigned to one haplotype. To prevent this, additional factors was added to the factor graph, each representing the probability mass function of a multinomial distribution with a number of categories equal to the number of haplotypes and with equal event probabilities. One factor vertex is constructed for each cut where the two haplotypes align; they encode our desire that at each place the models align, we wish the maps to be evenly distributed between haplotypes. The entire factor graph is presented in Figure 4.17.

Validation with Synthetic Data

As with the map refinement procedure, I tested the belief propagation-based haplotype separation approach extensively on a synthetic data set before applying it to real data. The experimental approach is similar: I began with the same reference genome restriction map, but constructed two "noisy" references from it. I used these noisy

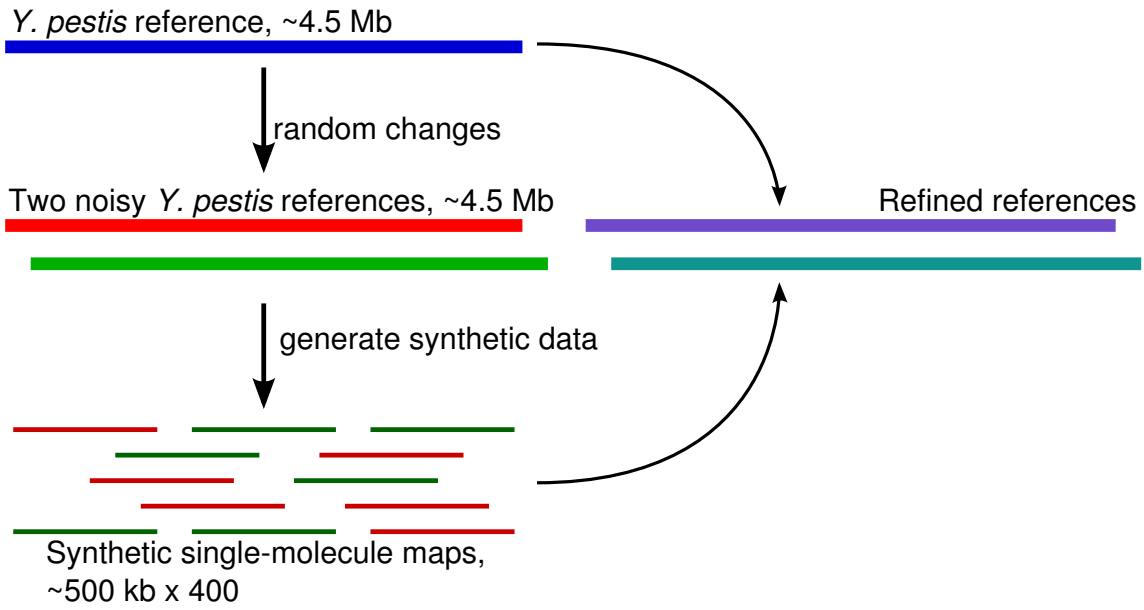

Figure 4.18. A strategy to use synthetic single-molecule maps to validate the belief propagation-based haplotype recovery algorithm.

references to create a combined synthetic data set of 400 synthetic restriction maps, then applied the haplotype-enabled refinement procedure to see if I could recover the two haplotypes represented by the original "noisy" references.

An example region is presented in Figure 4.19. The belief propagation approach was highly successful at recapitulating the two original "noisy" reference maps: for the first haplotype, 30 differences were recovered with only one false positive and one false negative. For the second haplotype, 23 separate differences were recovered, with two false positives and two false negatives.

Multiple Myeloma: Chromosome 13 Loss of Heterozygosity

In section 4.3, I noted that a loss of heterozygosity (LOH) on chromosome 13 in a multiple myeloma genome was associated with poor prognosis [43]. To assay the

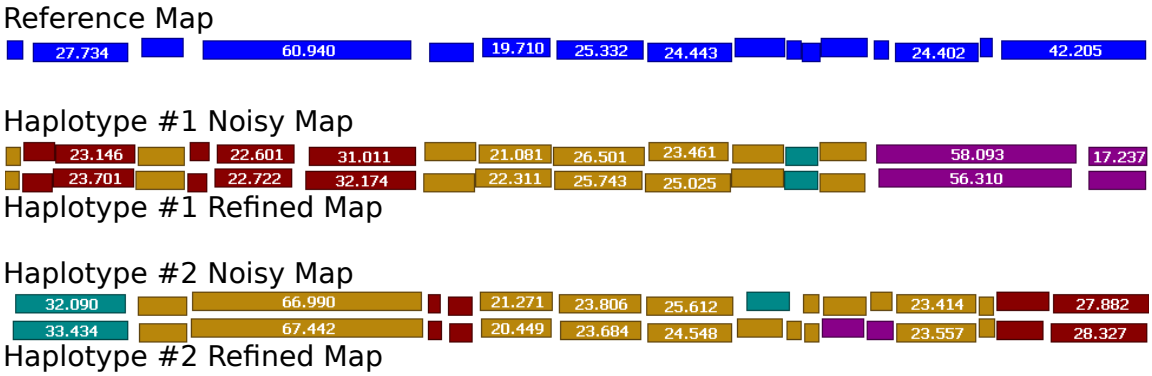

Figure 4.19. The belief propagation procedure correctly recovers synthetic haploid variation. The alignments show an example region from the *Y. pestis* genome, demonstrating the HMM- and belief propagation-based procedure recovers most of the "haploid" differences from a synthetic "diploid" data set.

possibility of LOH in Aditya Gupta's samples, I performed map refinement with belief propagation-based haplotype recovery on the normal (non-cancerous) genome and the post-chemotherapy resistance genome. The algorithm recovered 32 heterozygous sites on chromosome 13 from the normal sample, while it found only 3 heterozygous sites in the post-chemotherapy resistance sample. If in fact the entire chromosome had suffered a loss of heterozygosity, these could be due to contamination by normal cells or could simply be false positives. An example locus is presented in Figure 4.20. I also noted in section 4.3 that 22 of the 27 differences that were found by either the iterative assembly or the HMM but not by both appeared to be due to heterozygosity. Of these 22 ostensibly heterozygous loci, the belief propagation algorithm resolved 19 into two different haplotypes.

What might explain the widespread loss of heterozygosity on chromosome 13? A visual inspection of the post-resistance chromosome 13 assembly seemed to indicate substantially fewer single-molecule maps than aligned to other chromosomes, suggesting that a chromosome-wide deletion might be responsible. In fact, an Optical

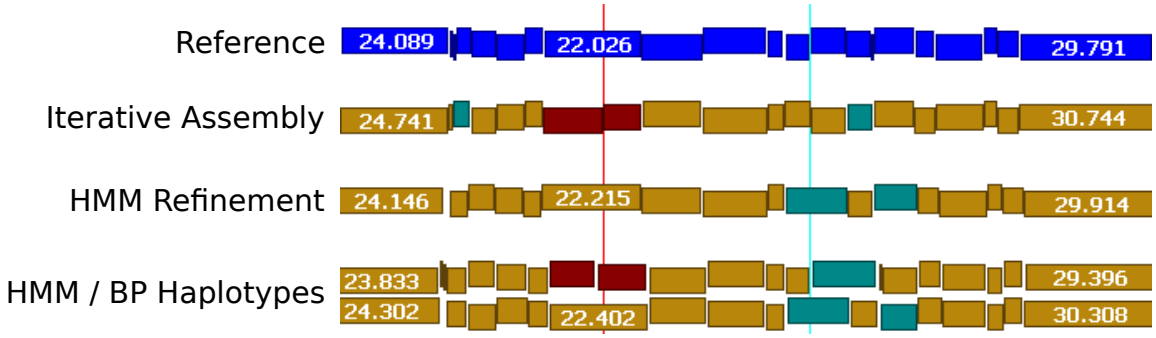

Figure 4.20. The belief propagation procedure recovers haploid variation from a human Optical Mapping data set. The alignments show an example region from chromosome 13 of the normal multiple myeloma data set, including two linked polymorphisms. The iterative assembly procedure recovered one of the haplotypes, while the HMM-based refinement procedure (without haplotypes) recovered the other. The HMM-based refinement procedure with belief propagation-based haplotype assignment recovered both.

Mapping coverage analysis based on an entirely different hidden Markov model [27] suggests that both the pre- and post-chemotherapy resistant multiple myeloma samples saw the deletion of an entire copy of chromosome 13, likely explaining the loss of heterozygosity that I observed.

Human Embryonic Stem Cells: Loss of Heterozygosity as a Culture Artifact

Human embryonic stem (hES) cells are pluripotent cells derived from human embryos [51]; they hold promise for greater understanding of tissue development and differentiation [52–54], as well potential therapeutic value [55–57]. Unfortunately, deriving human embryonic stem cells is fraught both technically [51] and legally [58]; but extended propagation of an existing hES cell line may induce genetic instability and loss of heterozygosity [59, 60] as subpopulations acquire favorable mutations and

	Heterozygous in P22	Homozygous in P22
Heterozygous in P208	48	159
Homozygous in P208	394	294

Table 4.2. The belief propagation procedure uncovers loss of heterozygosity as embryonic stem cell lines are passaged. Of the 895 variants found in chromosomes 1-5 of H1 P22 and H1 P208, 394 loci were heterozygous in P22 and homozygous in P208.

outcompete their neighbors.

In order to assess whether hES cell lines suffer loss of heterozygosity at a structural level as well as at the level of sequence, we created Optical Mapping data sets from the H1 hES cell line in collaboration with the laboratory of Jamie Tompson [61, 62]. We studied the cell line at two different passages: passage 22 (H1 P22) and passage 208 (H1 P208). I used the HMM-based haplotype recovery procedure to analyze the entirety of the H1 P22 genome and chromosomes one thru five of H1 P208. After manual curation, the H1 P22 genome yielded 1544 structural variants, 977 of which were heterozygous. In the first five chromosomes of the H1 P208 genome, 394 of these loci had become homozygous. (Also, 159 loci went from homozygous to heterozygous, likely indicating a mixture of false positives, false negatives and culture-induced mutation.) These results are tabulated in Table 4.2, and an example region is presented in Figure 4.21.

4.5 Conclusion

Hidden Markov models have become a *de rigueur* tool in any bioinformatician's toolbox. Many problems, especially in the realm of sequence analysis, can be cast as "labelling" problems [63]: which regions of this sequence code for proteins [14]?

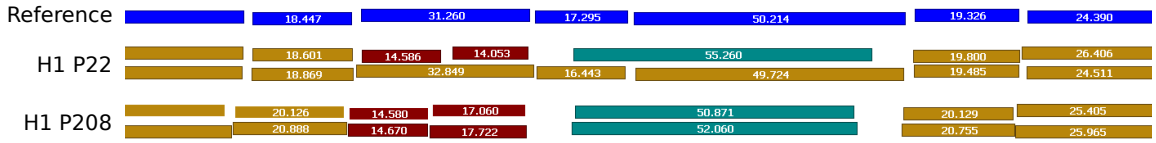

Figure 4.21. The belief propagation procedure uncovers loss of heterozygosity as embryonic stem cell lines are passaged. The alignments show an example region of chromosome 2 from the H1 P22 and H1 P208 genomes. In H1 P22, the region is heterozygous, with one haplotype evincing a substantial deletion in comparison to the reference. In H1 P208, the deletion is present in both haplotypes at this locus.

Contain CpG islands [15]? Preferentially bind histones [64]? The power of a hidden Markov model to answer these questions comes from its ability to reconnect the underlying state space (representing the physical process being studied) and the output space (the data observed). The evaluation, decoding and learning problems have algorithms that are efficient, effective, and theoretically well-founded.

The analyses we wish to perform on the data produced by the Optical Mapping platform frequently have at their core a labelling problem as well: which restriction fragment(s) in the genome produced each of the fragments we observe in the data set? The answer to this question underlies algorithms that use the HMM to do reference map discrimination, pairwise alignment and parameter estimation. For pairwise alignment and reference map discrimination, effective tools already exist, allowing us to validate the HMM-based methods and the model on which they're based. For parameter estimation, however, the hidden Markov model and algorithms represent an important addition to the Optical Mapping analysis toolbox.

Beyond the well-known and widely used solutions to the evaluation, decoding and learning problems, hidden Markov models provide a probabilistic foundation on which to build more tightly domain-coupled analyses. The reference refinement and

haplotype discernment algorithms are good examples: the former is a careful search through model space, informed less by general HMM theory than by the particular structure of the problem at hand. On the other hand, the haplotype discernment algorithm uses the HMM as an integral component in a larger belief propagation framework that performs inference on both the model and the data set simultaneously. In this way, the belief propagation approach can leverage the efficiency and sensitivity of the HMM-based methods to solve a massive optimization problem efficiently and effectively.

More generally, the success of these two algorithms demonstrates the power of a Bayesian approach to Optical Mapping analyses in particular and bioinformatics in general. When analyzing noisy data sets, a frequentist approach based on statistical tests may not have sufficient power to overcome the noise, especially in light of the usual need for multiple test correction [65–67]. A Bayesian approach, especially one based on model selection, may have better success [68, 69], especially in cases where prior belief does not favor one model strongly over others (i.e. there's no obvious "null hypothesis.") The most common practical criticism, that Bayesian methods are too computationally intensive for day-to-day use, is mitigated both by clever (domain-specific) ways to select candidate models, and by the ever-growing computational resources resulting from inexorable march of Moore's Law.

The ensembles of single-molecule observations produced by the Optical Mapping platform, suitably analyzed, are unique in their ability to provide insight into genome structure. The hidden Markov model described in this chapter represents a new formalism with which to build new tools for wringing biological meaning from these data sets and bringing additional power to this high-throughput single-molecule system for genome analysis.

4.6 Bibliography

[1] Nemethy, G. and Gibson, K. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry*, **96**:6472–6484, 1992.

[2] Kaźmierkiewicz, R., Liwo, A., and Scheraga, H. A. Addition of side chains to a known backbone with defined side-chain centroids. *Biophysical chemistry*, **100**(1-3):261–80, 2003.

[3] The International Phonetic Association. *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[4] Strang, G. *Introduction to Linear Algebra*. Wellesley Cambridge Press, 4th edition, 2009.

[5] Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. 1993.

[6] Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, 1989.

[7] Broman, K. W., et al. R/qtl: QTL mapping in experimental crosses. *Bioinformatics (Oxford, England)*, **19**(7):889–90, 2003.

[8] Collier, N., Nobata, C., and Tsujii, J. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of COLING*, pages 201–207. Saarbruecken, 2000.

[9] Mitchison, G. J. A probabilistic treatment of phylogeny and sequence alignment. *Journal of molecular evolution*, **49**(1):11–22, 1999.

[10] Shah, S. P., et al. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics (Oxford, England)*, **22**(14):e431–9, 2006.

[11] Colella, S., et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research*, **35**(6):2013–25, 2007.

[12] Wang, K., et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*, **17**(11):1665–74, 2007.

[13] Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome informatics International Conference on Genome Informatics*, **23**(1):205–11, 2009.

[14] Lukashin, A. V. and Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic acids research*, **26**(4):1107–15, 1998.

[15] Zhang, X., et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, **126**(6):1189–201, 2006.

[16] Reese, M. G., et al. Improved splice site detection in Genie. *Journal of computational biology : a journal of computational molecular cell biology*, **4**(3):311–23, 1997.

[17] Asai, K., Hayamizu, S., and Handa, K. Prediction of protein secondary structure by the hidden Markov model. *Computer applications in the biosciences : CABIOS*, **9**(2):141–6, 1993.

[18] Cole, C., Barber, J. D., and Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic acids research*, **36**(Web Server issue):W197–201, 2008.

[19] Karplus, K., Barrett, C., and Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics (Oxford, England)*, **14**(10):846–56, 1998.

[20] Valouev, A., et al. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, **22**(10):1217–24, 2006.

[21] Dimalanta, E. T., et al. A microfluidic system for large DNA molecule arrays. *Analytical chemistry*, **76**(18):5293–301, 2004.

[22] Lander, E. S. and Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3):231–9, 1988.

[23] Anantharaman, T. S., Mishra, B., and Schwartz, D. C. Genomics via Optical Mapping III: Contiging Genomic DNA and Variations. Technical Report TR1998-760, New York University, 1998.

[24] Karp, R. M. and Shamir, R. Algorithms for optical mapping. In *RECOMB*, pages 117–124. 1998.

[25] Lee, J. K., Dancík, V., and Waterman, M. S. Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. In *RECOMB*, pages 147–152. 1998.

[26] Valouev, A., et al. Alignment of optical maps. *Journal of computational biology : a journal of computational molecular cell biology*, **13**(2):442–62, 2006.

[27] Sarkar, D. *On the Analysis of Optical Mapping Data*. Ph.D. thesis, University of Wisconsin – Madison, 2006.

[28] Sarkar, D., et al. Statistical significance of optical map alignments. *Journal of computational biology : a journal of computational molecular cell biology*, **19**(5):478–92, 2012.

[29] Durbin, R. M., et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[30] Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2):260–269, 1967.

[31] Baum, L. E., et al. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics1*, **41**(1):164–171, 1970.

[32] Teague, B., et al. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(24):10,848–53, 2010.

[33] Anantharaman, T., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. III: Contiging genomic DNA. In *Proc Int Conf Intell Syst Mol Biol*, Proceedings 7th Intl. Cnf. on Intelligent Systems for Molecular Biology, pages 18–27. 1999.

[34] Litzkow, M., Livny, M., and Mutka, M. Condor - A Hunter of Idle Workstations.

In *Proceedings of the 8th International Conference of Distributed Computing Systems*. 1988.

[35] Tannenbaum, T., et al. Condor - A Distributed Job Scheduler. In T. Sterling (editor), *Beowulf Cluster Computing with Linux*. MIT Press, 2001.

[36] Anantharaman, T. S., Mishra, B., and Schwartz, D. C. Genomics via optical mapping. II: Ordered restriction maps. *Journal of computational biology : a journal of computational molecular cell biology*, **4**(2):91–118, 1997.

[37] Vetterling, W. T. and Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, 2007.

[38] Hirotugu, A. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6):716–723, 1974.

[39] Schwarz, G. E. Estimating the dimension of a model. *Annals of Statistics*, **6**(2):461–464, 1978.

[40] Link, W. A. and Barker, R. J. Model weights and the foundations of multimodel inference. *Ecology*, **87**(10):2626–35, 2006.

[41] Kyle, R. A. and Rajkumar, S. V. Multiple myeloma. *Blood*, **111**(6):2962–72, 2008.

[42] Child, J. A., et al. High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma. *The New England journal of medicine*, **348**(19):1875–83, 2003.

[43] Tricot, G., et al. Poor prognosis in multiple myeloma is associated only with partial or complete deletions of chromosome 13 or abnormalities involving 11q and not with other karyotype abnormalities. *Blood*, **86**(11):4250–6, 1995.

[44] Fujita, P. A., et al. The UCSC Genome Browser database: update 2011. *Nucleic acids research*, **39**(Database issue):D876–82, 2011.

[45] Hardisty, E. and Resnik, P. Gibbs Sampling for the Uninitiated. *Bernoulli*, **4956**(June), 2010.

[46] Pearl, J. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence*, pages 133–136. AAAI Press, Menlo Park, CA, 1982.

[47] Yedidia, J. S., Freeman, W. T., and Weiss, Y. Understanding Belief Propagation and its Generalizations. Technical report, Mitsubishi Electric Research Laboratories, 2001.

[48] Kschischang, F., Frey, B., and Loeliger, H.-A. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, **47**(2):498–519, 2001.

[49] Ihler, A. T., Fisher, J. W., and Willsky, A. S. Loopy Belief Propagation : Convergence and Effects of Message Errors. *Journal of Machine Learning Research*, **6**:905–936, 2005.

[50] Yedidia, J. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Advances in neural information …*, 2001.

[51] Thomson, J. A., et al. Embryonic stem cell lines derived from human blastocysts. *Science (New York, NY)*, **282**(5391):1145–7, 1998.

[52] Amit, M., et al. Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Developmental biology*, **227**(2):271–8, 2000.

[53] Pan, G. and Thomson, J. A. Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell research*, **17**(1):42–9, 2007.

[54] Ying, Q.-L., et al. The ground state of embryonic stem cell self-renewal. *Nature*, **453**(7194):519–23, 2008.

[55] Aiuti, A., et al. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science (New York, NY)*, **296**(5577):2410–3, 2002.

[56] Silani, V., et al. Stem-cell therapy for amyotrophic lateral sclerosis. *Lancet*, **364**(9429):200–2, 2004.

[57] Segers, V. F. M. and Lee, R. T. Stem-cell therapy for cardiac disease. *Nature*, **451**(7181):937–42, 2008.

[58] Kaiser, J. A Legal Win for Stem Cell Research, but Case May Not Be Over. *Science Insider*, 2012.

[59] Lefebvre, L., et al. Selection for transgene homozygosity in embryonic stem cells results in extensive loss of heterozygosity. *Nature genetics*, **27**(3):257–8, 2001.

[60] Närvä, E., et al. High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nature biotechnology*, **28**(4):371–7, 2010.

[61] Ananiev, G. E., et al. Optical mapping discerns genome wide DNA methylation profiles. *BMC molecular biology*, **9**:68, 2008.

[62] Ananiev, G. E. *Structural and epigenetic analysis of the human genome.* Ph.D. thesis, 2008.

[63] Eddy, S. R. What is a hidden Markov model? *Nature biotechnology*, **22**(10):1315–6, 2004.

[64] Xu, H., et al. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics (Oxford, England)*, **24**(20):2344–9, 2008.

[65] Duggal, P., et al. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC genomics*, **9**:516, 2008.

[66] Noble, W. S. How does multiple testing correction work? *Nature biotechnology*, **27**(12):1135–7, 2009.

[67] van Iterson, M., Boer, J. M., and Menezes, R. X. Filtering, FDR and power. *BMC bioinformatics*, **11**:450, 2010.

[68] Shoemaker, J. S., Painter, I. S., and Weir, B. S. Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics*, **15**(9):354–358, 1999.

[69] Beaumont, M. a. and Rannala, B. The Bayesian revolution in genetics. *Nature reviews Genetics*, **5**(4):251–61, 2004.

5 CONCLUSION

The work described in this thesis was aimed at advancing our understanding of normal structural polymorphism in the human genome. This brief conclusion describes the next steps for the field of medical genetics in general and for Optical Mapping in particular.

5.1 The Future of Personalized Medicine: From Risk Factors to Tailored Treatments

The popular press greets the results of each new high-profile association or case-control study by announcing that scientists have "found the gene for" whatever common disease or trait the study was targetting [1–3]. Of course this is very rarely the case [4, 5]; in most cases [6–8], the study has identified variants associated with the disease or trait in question, or more infrequently [9–11] the alleles responsible for a Mendelian sub-phenotype. This work has been progressing since well before the sequencing of the human genome [12–14] and while genetic testing for Mendelian genetic disorders is well-established [15], only a handful of genetic variants related to common diseases are commonly used to direct treatment. Perhaps the best known (and certainly the most litiginous [16]) are the genetic tests for mutations in BRCA1 and BRCA2, variants that have substantial impact on a woman's risk of developing breast cancer [17–19] and have become common in directing the course of treatment for that disease [20, 21]. Another more recent example is the impact of variants in the CYP2C9 and CYP4F2 genes, whose impact on coagulation [22, 23] is being used to direct Warfarin anticoagulant therapy [24].

Despite these limited inroads, the technology to sequence a patient's exome [25–28] or their entire genome [29–31] is slowing making its way into the clinic, threatening to make locus-specific testing obsolete [32]. What stands in the way of sequencing having more of an clinical impact? The most obvious is our limited understanding of the *functional impact* of the variants we observe [33]. Even for exonic single-nucleotide variants that are predicted to result in the protein's loss of function, our understanding of the underlying biological networks is so limited that we cannot *a priori* assign a phenotypic outcome [34, 35]. What then are we to do with variants in the regulatory and non-coding regions of the genome? Protein-coding open reading frames account for approximately 1.5% of the human genome, but analyses of genome evolution find 10-fold more basepairs are conserved (and thus putatively functional) than are present in protein-coding open reading frames [36]; and recent ENCODE results [37] assign a biochemical or regulatory function to an whopping 80% of the genome, the functional significance of which remains poorly understood.

The problem is exacerbated by recent evidence that a large proportion, if not a majority, of any individual's genetic polymorphism is composed of rare variants [38, 39]. This has been shown true for both single-nucleotide variants in exons [40] and for copy number polymorphisms [41]. Rare variants are problematic because a genome-wide association or case-control study to probe their phenotypic effects would be prohibitively large [38], and these approaches are currently our most powerful tools for connecting genotype to phenotype in a hypothesis-free manner [33]. Fortunately, common alleles exist with which to infer the function of conserved genomic elements [38]; but predicting how rare polymorphisms alter those functions is as of yet beyond our reach.

Finally, even with full knowledge of an individual's genotype, that person's

genetic makeup does not fully explain his or her propensity to develop common diseases [42–47]: for most, heritability is estimated at less than 50%. This can be partly explained by traditional Darwinian selection: alleles that are deleterious to reproduction are subject to strong negative selective pressure and are maintained at only a low frequency in the population [48]. In fact, the diseases that are most strongly heritable (like age-related macular degeneration [46] and Alzheimer disease[47]) only manifest themselves after an individual's reproductive prime. Other alleles are subject to balancing selection, where a phenotype that might be considered deleterious in one environment is beneficial in another. The most famous example is the deletion which causes α-thalassemia (sickle cell disease) when homozygous, but confers resistance to malaria infection in heterozygotes and is thus maintained at a relatively high frequency in populations where malaria is common [49]. There is also growing evidence for balancing selection in alleles of lesser penetrance, including a polymorphism which seems to confer higher risk of developing diabetes but lower risk of inflammatory bowel disease [50].

These explanations all stem from classical Mendelian genetics, but multi-locus traits [51] clearly play an important role as well. Epistasis and allele interaction must be considered [52–54] while epigenetic effects remain largly unexplored [55, 56]. Finally, the interaction of genetics and the environment, nature and nurture, limit the use of even the most comprehensive, explanatory genomic analysis [57–60].

All this goes to show that in translating the biological revolution potentiated by the human genome sequence into the clinic, genomics in isolation will not be enough. The human genome sequence has provided a scaffold with which to study the rest of human biology in a comprehensive manner, free from the constraints of hypothesis-driven scientific approaches [61]. Not gross disease states but measures of

transcripts, proteins, metabolites, lipids, even an individual's microbiome – all serve as phenotypic expressions of genotype. These measurements can provide information about a person's health at a much more granular, detailed level than the traditional distinction between "health" and "disease" [62, 63]. Comprehensive, multi-scale, longitudinal records [64, 65] of the biological systems and networks that keep us healthy will contribute to our understanding of how their dysregulation results in common diseases, diseases that make up so much of the world's modern healthcare burden. Their application in the clinic, in the hands of skilled diagnosticians, has the potential to deliver on the human genome project's as-of-yet unfulfilled promise of personalized medicine.

5.2 Cancer: The High Ground

Such "personalized medicine" approaches will find their true test, and perhaps their greatest success, in the treatment of cancer. Because each malignancy arises *de novo*, each is a unique disease with a unique (set of) genome(s). This heterogeneity in genotype and phenotype is the root cause of the hetergeneity in cancer therapy success [66]. Understanding each tumor's unique genetic makeup [30, 67] and its regulatory [68–70], proteomic [71] and metabolomic [72–74] profiles could inform a unique treatment regime, tailored to specifically disrupt that tumor with minimum damage to healthy tissues. Understanding the extreme dysregulation of a malignancy's normal cellular processes will depend strongly on our understanding of those processes' and networks' normal modes of function [75–77].

5.3 Optical Mapping and its Successors

Most analyses of the human genome (and many other global methods such as proteomics and transcriptomics) rely heavily on the reference genome sequence to anchor the short-range data they produce: short sequence reads [78, 79], peptide fragments [80, 81], etc. The Optical Mapping platform's approach to the manipulation, presentation and analysis of large, single DNA molecules promises a direct route to the long-range information that contemporary modes of genome analysis lack. For example, as discussed in Chapter 1, our understanding of normal germ-line polymorphic variation in genome structure remains incomplete [82]. It's not yet clear whether structural variants are predominantly rare, as seems to be the case with single-nucleotide variants. Forty-four percent of the autosomal copy number variants found in the Wellcome Trust Case-Control Consortium's study of over 10,000 genomes [41] were found with a minor allele frequency of less than 5%, but it's not clear whether the processes that give rise to copy number variants (and genome structure variation in general) [83] would increase the likelihood of the same variant recurring at the same spot [84] (making them more frequent and easier to genotype), or would be responsible for an outsized rate of *de novo* occurrence and a preponderance of rare alleles [85].

The most exiciting opportunities for Optical Mapping, however, lie in sorting out somatic mutations: that is, in studying cancer. The most powerful analyses to understand a tumor's genotype (and resultant phenotype) would be based on a fully sequenced, assembled genome. However, because tumor genomes are amplified and rearranged [86, 87] almost by definition, long-range information is required to assemble the small reads produced by modern sequencing methods.

What advances are required to enable this kind of analysis? First and foremost, improved data throughput and data quality are necessary. Advances in DNA enzymology [88], molecule presentation [89, 90] and automated data collection and analysis may soon reduce the data collection time for a human (or human-derived) genome from weeks to hours. Next, new computational approaches will be required to deal with this deluge of data. There's no getting around it: genome assembly is a global optimization problem [91–95] and will require global approaches akin to the belief propagation methods presented in Chapter 4. However, cutting-edge techniques in computer science [96, 97] can reduce the computational effort by orders of magnitude. So can the move away from computation on general-purpose processors to more specialized hardware such as graphics processing units [98] or field-programmable gate arrays [99]. Finally, in the absence of any of these advances, Moore's Law marches ever on, increasing both our ability to store the data and our facility in analyzing it.

5.4 Conclusion

In a volume published in 1917 [100], Carl Wolfgang Ostwald described the world between the structures visible under the optical microscope and the molecules studied by chemists "the world of the ignored dimensions." For decades, our understanding of the genome was limited to the structural changes we could see in a microscope, and the molecular changes we found using sequencing methods. As Optical Mapping probes into the "ignored dimensions" between these two extremes, I am reminded of zooming into a Mandelbrot set [101] fractal: instead of simple, definite answers to our questions, we uncover more complexity and more questions to ask.

Why bother study the human genome? We do so for the promise of medical advance

potentiated by biological insight. Eric Lander states that our goal is "to transform the treatment of common disease through an understanding of the underlying molecular pathways." [33] The human genome is the scaffold with which we contextualize the rest of human biology, on scales that reach from cells to populations and from milliseconds to eons. This thesis advances our understanding of the human genome, and the approaches embodied herein and the artifacts produced as a result lay the groundwork for Optical Mapping to continue to expand our knowledge of its intricacies.

5.5 Bibliography

[1] Type 2 diabetes genes mapped out, 2007.

[2] Obesity gene 'affects appetite', 2008.

[3] Hesman, T. Three Genes Linked to Alzheimer's Disease Risk, 2009.

[4] Neale, B. M., et al. Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proceedings of the National Academy of Sciences of the United States of America*, **107**(16):7395–400, 2010.

[5] Chen, W., et al. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(16):7401–6, 2010.

[6] Easton, D. F., et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**(7148):1087–93, 2007.

[7] McPherson, R., et al. A common allele on chromosome 9 associated with coronary heart disease. *Science (New York, NY)*, **316**(5830):1488–91, 2007.

[8] Sladek, R., et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**(7130):881–5, 2007.

[9] Liu, X.-Q., Paterson, A. D., and Szatmari, P. Genome-wide linkage analyses of quantitative and categorical autism subphenotypes. *Biological psychiatry*, **64**(7):561–70, 2008.

[10] Jugessur, A., et al. Fetal genetic risk of isolated cleft lip only versus isolated cleft lip and palate: a subphenotype analysis using two population-based studies of orofacial clefts in Scandinavia. *Birth defects research Part A, Clinical and molecular teratology*, **91**(2):85–92, 2011.

[11] Fischer, A., et al. Association of inflammatory bowel disease risk loci with sarcoidosis, and its acute and chronic subphenotypes. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*, **37**(3):610–6, 2011.

[12] Wooster, R., et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**(6559):789–92, 1995.

[13] Lifton, R. P. Molecular genetics of human blood pressure variation. *Science (New York, NY)*, **272**(5262):676–80, 1996.

[14] Yamagata, K., et al. Mutations in the hepatocyte nuclear factor-1alpha gene in maturity-onset diabetes of the young (MODY3). *Nature*, **384**(6608):455–8, 1996.

[15] Dolan, S. M. Prenatal genetic testing. *Pediatric annals*, **38**(8):426–30, 2009.

[16] Marshall, E. Myriad Genetics Wins and Loses in Latest Court Ruling, 2012.

[17] Ford, D., et al. Risks of cancer in BRCA1-mutation carriers. *Lancet*, **343**(8899):692–5, 1994.

[18] Wooster, R., et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science (New York, NY)*, **265**(5181):2088–90, 1994.

[19] Easton, D. F., et al. Breast and ovarian cancer incidence in BRCA1-mutation carriers. *American journal of human genetics*, **56**(1):265–71, 1995.

[20] Schrag, D., et al. Life expectancy gains from cancer prevention strategies for women with breast cancer and BRCA1 or BRCA2 mutations. *JAMA : the journal of the American Medical Association*, **283**(5):617–24, 2000.

[21] Sandhaus, L. M., et al. Reporting BRCA test results to primary care physicians. *Genetics in medicine : official journal of the American College of Medical Genetics*, **3**(5):327–34, 2001.

[22] Steward, D. J., et al. Genetic association between sensitivity to warfarin and expression of CYP2C9*3. *Pharmacogenetics*, **7**(5):361–7, 1997.

[23] Caldwell, M. D., et al. CYP4F2 genetic variant alters required warfarin dose. *Blood*, **111**(8):4106–12, 2008.

[24] Frueh, F. W. On rat poison and human medicines: personalizing warfarin therapy. *Trends in molecular medicine*, **18**(4):201–5, 2012.

[25] Ng, S. B., et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, **42**(1):30–5, 2010.

[26] Ng, S. B., et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics*, **42**(9):790–3, 2010.

[27] O'Roak, B. J., et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics*, **43**(6):585–9, 2011.

[28] Dixon-Salazar, T. J., et al. Exome sequencing can improve diagnosis and alter patient management. *Science translational medicine*, **4**(138):138ra78, 2012.

[29] Bainbridge, M. N., et al. Whole-genome sequencing for optimized patient management. *Science translational medicine*, **3**(87):87re3, 2011.

[30] Welch, J. S., et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA : the journal of the American Medical Association*, **305**(15):1577–84, 2011.

[31] Hayden, E. C. Sequencing set to alter clinical landscape. *Nature*, **482**(7385):288, 2012.

[32] Drmanac, R. Medicine. The ultimate genetic test. *Science (New York, NY)*, **336**(6085):1110–2, 2012.

[33] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature*, **470**(7333):187–97, 2011.

[34] Roberts, N. J., et al. The predictive capacity of personal genome sequencing. *Science translational medicine*, **4**(133):133ra58, 2012.

[35] MacArthur, D. G., et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, **335**(6070):823–828, 2012.

[36] Lowe, C. B., et al. Three periods of regulatory innovation during vertebrate evolution. *Science (New York, NY)*, **333**(6045):1019–24, 2011.

[37] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74, 2012.

[38] McClellan, J. and King, M.-C. Genetic heterogeneity in human disease. *Cell*, **141**(2):210–7, 2010.

[39] 1000 Genomes Project Consortium, et al. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–73, 2010.

[40] Nelson, M. R., et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, NY)*, **337**(6090):100–4, 2012.

[41] Craddock, N., et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**(7289):713–20, 2010.

[42] Schildkraut, J. M., Risch, N., and Thompson, W. D. Evaluating genetic association among ovarian, breast, and endometrial cancer: evidence for a breast/ovarian cancer relationship. *American journal of human genetics*, **45**(4):521–9, 1989.

[43] Poulsen, P., et al. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, **42**(2):139–45, 1999.

[44] Katzmarzyk, P. T., et al. Familial resemblance for coronary heart disease risk: the HERITAGE Family Study. *Ethnicity & disease*, **10**(2):138–47, 2000.

[45] Grönberg, H. Prostate cancer epidemiology. *Lancet*, **361**(9360):859–64, 2003.

[46] Seddon, J. M., et al. The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Archives of ophthalmology*, **123**(3):321–7, 2005.

[47] Gatz, M., et al. Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry*, **63**(2):168–74, 2006.

[48] Hardy, G. H. Mendelian proportions in a mixed population. *Science (New York, NY)*, **28**(706):49–50, 1908.

[49] Allison, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *British medical journal*, **1**(4857):290–4, 1954.

[50] Wang, K., et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Human molecular genetics*, **19**(10):2059–67, 2010.

[51] Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, **52**:399–433, 1918.

[52] Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, **56**(1-3):73–82, 2003.

[53] Carlson, C. S., et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American journal of human genetics*, **74**(1):106–20, 2004.

[54] Hirschhorn, J. N. and Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature reviews Genetics*, **6**(2):95–108, 2005.

[55] Fraga, M. F., et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(30):10,604–9, 2005.

[56] Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**(7143):433–40, 2007.

[57] Freeman, G. H. Statistical methods for the analysis of genotype-environment interactions2. *Heredity*, **31**(3):339–354, 1973.

[58] Via, S. and Lande, R. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*, **39**(3):505–522, 1985.

[59] Westcott, B. Some methods of analysing genotypeâŁ"environment interaction. *Heredity*, **56**(2):243–253, 1986.

[60] Gillespie, J. H. and Turelli, M. Genotype-environment interactions and the maintenance of polygenic variation. *Genetics*, **121**(1):129–38, 1989.

[61] Mill, J. S. *A System of Logic*. Harper & Brothers, New York, 8th edition, 1882.

[62] Blau, C. A. Can intensive longitudinal monitoring of individuals advance cancer research? *The oncologist*, **17**(5):587–9, 2012.

[63] Chen, R., et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, **148**(6):1293–307, 2012.

[64] Shah, N. H. and Tenenbaum, J. D. The coming age of data-driven medicine: translational bioinformatics' next frontier. *Journal of the American Medical Informatics Association : JAMIA*, **19**(e1):e2–e4, 2012.

[65] Smarr, L. Quantifying your body: A how-to guide from a systems biology perspective. *Biotechnology journal*, **7**(8):980–91, 2012.

[66] Fidler, I. J. Tumor heterogeneity and the biology of cancer invasion and metastasis. *Cancer research*, **38**(9):2651–60, 1978.

[67] Roychowdhury, S., et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science translational medicine*, **3**(111):111ra121, 2011.

[68] Rhodes, D. R. and Chinnaiyan, A. M. Integrative analysis of the cancer transcriptome. *Nature genetics*, **37 Suppl**:S31–7, 2005.

[69] Maher, C. A., et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**(7234):97–101, 2009.

[70] Smith, T. M., Olson, N., and Smith, D. Making cancer transcriptome sequencing assays practical for the research and clinical scientist. *Genome Biology*, **11**(Suppl 1):P39, 2010.

[71] Hanash, S. and Taguchi, A. The grand challenge to decipher the cancer proteome. *Nature reviews Cancer*, **10**(9):652–60, 2010.

[72] Abate-Shen, C. and Shen, M. M. Diagnostics: The prostate-cancer metabolome. *Nature*, **457**(7231):799–800, 2009.

[73] McCarthy, N. Systems biology: Lethal weaknesses. *Nature reviews Cancer*, **11**(8):538–9, 2011.

[74] Clyne, M. Kidney cancer: Metabolomics for targeted therapy. *Nature reviews Urology*, **9**(7):355, 2012.

[75] Huang, S., Ernberg, I., and Kauffman, S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in cell & developmental biology*, **20**(7):869–76, 2009.

[76] Kreeger, P. K. and Lauffenburger, D. A. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, **31**(1):2–8, 2010.

[77] Roukos, D. H. Novel clinico-genome network modeling for revolutionizing genotype-phenotype-based personalized cancer care. *Expert review of molecular diagnostics*, **10**(1):33–48, 2010.

[78] Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, **18**(11):1851–8, 2008.

[79] Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14):1754–60, 2009.

[80] Washburn, M. P., Wolters, D., and Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology*, **19**(3):242–7, 2001.

[81] Sadygov, R. G., Cociorva, D., and Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods*, **1**(3):195–202, 2004.

[82] Alkan, C., Coe, B. P., and Eichler, E. E. Genome structural variation discovery and genotyping. *Nature reviews Genetics*, **12**(5):363–76, 2011.

[83] Bailey, J. A. and Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews Genetics*, **7**(7):552–64, 2006.

[84] Turner, D. J., et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature genetics*, **40**(1):90–5, 2008.

[85] Stankiewicz, P. and Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual review of medicine*, **61**:437–55, 2010.

[86] Kinzler, K. W. and Vogelstein, B. Lessons from Hereditary Colorectal Cancer. *Cell*, **87**(2):159–170, 1996.

[87] Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell*, **144**(5):646–74, 2011.

[88] Jo, K., Schramm, T. M., and Schwartz, D. C. A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding. *Methods in molecular biology (Clifton, NJ)*, **544**:29–42, 2009.

[89] Jo, K., et al. A single-molecule barcoding system using nanoslits for DNA analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(8):2673–8, 2007.

[90] Kounovsky-Shafer, K. L., et al. Presentation of large DNA molecules for analysis as nanoconfined dumbbells. *Proc Natl Acad Sci U S A*, **Under revi**, 2012.

[91] Pevzner, P. A., Tang, H., and Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(17):9748–53, 2001.

[92] Zhang, Y. and Waterman, M. S. An Eulerian path approach to global multiple alignment for DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology*, **10**(6):803–19, 2003.

[93] Butler, J., et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, **18**(5):810–20, 2008.

[94] Medvedev, P. and Brudno, M. Maximum likelihood genome assembly. *Journal of computational biology : a journal of computational molecular cell biology*, **16**(8):1101–16, 2009.

[95] Miller, J. R., Koren, S., and Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics*, **95**(6):315–27, 2010.

[96] Gonzalez, J. E., Low, Y., and Guestrin, C. Residual Splash for Optimally Parallelizing Belief Propagation. *Artificial Intelligence*, **5**:177–184, 2009.

[97] Gonzalez, J., et al. Parallel gibbs sampling: From colored fields to thin junction trees. In *In Artificial Intelligence and Statistics (AISTATS)*, volume 15. 2011.

[98] Charalambous, M., Trancoso, P., and Stamatakis, A. Initial Experiences Porting a Bioinformatics Application to a Graphics Processor. *Lecture Notes in Computer Science*, **3746**:415–425, 2005.

[99] Oliver, T., et al. Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics (Oxford, England)*, **21**(16):3431–2, 2005.

[100] Ostwald, C. W. W. *An Introduction to Theoretical and Applied Colloid Chemistry.* Wiley, 1917.

[101] Branner, B. The Mandelbrot set. In *Proceedings of Symposia in Applied Mathematics*, pages 65–85. 1989.

DISCARD THIS PAGE

COLOPHON

This thesis was produced entirely using Free and Open Source Software.

- R, **http://www.r-project.org**, for graphs and plots.
- Inkscape, **http://www.inkscape.org**, for vector graphics.
- The GNU Image Manipulation Program, **http://www.gimp.org**, for raster graphics.
- LaTeX, for typesetting.

The text is typeset in Gyre Pagella font, while equations remain typeset in TeX's default font (Computer Modern.) Based on Will Benton's UW – Madison thesis style, as modified by Steven Baumgart.