OPTICAL MAPPING AND DNA SEQUENCING APPROACHES REVEAL WIDESPREAD STRUCTURAL AND COPY NUMBER VARIATION IN MULTIPLE MYELOMA

By

Aditya Gupta

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biophysics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2015

Date of final oral examination: January 13th, 2015

The dissertation is approved by the following members of the Final Oral Committee:
David C. Schwartz, Professor, Chemistry and Genetics
Fotios Asimakopoulos, Assistant Professor, Medicine
Aseem Z. Ansari, Professor, Biochemistry
Michael N. Gould, Professor, Oncology
Douglas B. Weibel, Associate Professor, Biochemistry, Chemistry and Biomedical Engineering

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Prof. David C. Schwartz, for accepting me into his lab all those years ago and for being a constant pillar of inspiration and support over the years. I am also thankful to Prof. Fotis Asimakopoulos for being a wonderful collaborator and for his guidance during my research work. Also, I would like to thank the rest of my committee for their attention to and support of my research work.

I am grateful to the members of Schwartz lab for their continuous support and for engaging in conversations that helped me shape my research. Special thanks to Michael Place for helping me out whenever I needed any kind of help and for listening to me patiently. Thanks to Steve Goldstein and Gus Potamousis for their extensive support and for all that I have learned from them. Thanks to Kristy Kounovsky-Shafer for introducing me to Adobe Illustrator and to Mohana Ray for her guidance all along.

I also thank Dr. Louise Pape and the Biophysics Graduate Program staff, especially Dr. Kim Voss and Katherine Rankin McVey for their continued support over the years.

When I first moved to Madison from India for graduate school, I wasn't sure how the transition to a new country would feel like. I feel extremely fortunate to have met many good friends here who have made this place feel like home, and for that I thank all of them.

Finally, I would like to extend my deepest gratitude to my family for their unconditional love and support all my life. I dedicate this thesis to my parents, without whose love and guidance none of this would have ever happened.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Cancer genomes are very complex and present an extreme range of mutational classes that challenge our abilities to comprehensively discover and understand them. Over the last five decades, cytogenetic approaches to this problem, and more recently, modern DNA sequencing approaches, have significantly advanced our understanding of cancer genome variation. However, our knowledge of genomic structural variation is attenuated due to the limitations inherent to these methods. In this thesis, we present the application of Optical Mapping, a single-molecule whole-genome analysis system, to address this shortcoming and study structural variation in a cancer genome. Through our analysis, we have identified widespread structural variation, copy number variation and genomic rearrangements in a multiple myeloma genome. Additionally, we describe our efforts towards comprehensive identification of genome structure and variation in this cancer genome by integrating findings from Optical Mapping and DNA sequencing based genomic analysis. Finally, by studying this multiple myeloma genome at two time-points during pathogenesis using aforementioned approaches, we have demonstrated an increase in mutational complexity with tumor progression at all length scales of variation. Overall, this thesis presents widespread structural variation in a cancer genome, identifies novel genomic targets that might play a role in tumor pathogenesis and drug resistance mechanisms, and highlights the need to routinely incorporate structural variation analysis while studying cancer genomes.

Chapter 1: Introduction

Cancer genomes harbor many somatic changes that range in size from single nucleotide to entire chromosomes. In this chapter, I will discuss the methods used to analyze cancer genomes, with a special emphasis on studying structural variation. I will also discuss the limitations of these methods. Finally, I have reviewed recent progress towards the understanding of genomic changes in multiple myeloma, highlighting our lack of knowledge about structural variation in multiple myeloma. In the following chapters, I have presented an integrative approach that uses Optical Mapping and DNA sequencing to characterize structural and other variation in multiple myeloma.

1.1. Cancer genome – history and development

Cancer has long been established as a disease of the genome, arising from a clone of cells that have acquired mutations leading to unregulated growth of these cells (Mitelman et al. 2004; Mitelman et al. 2007). While the physical description of patient symptoms led to the words *karkinos* (crab-like) and *oncos* (swelling), which were first used to describe cancers, it was the somatic mutation theory postulated by the German scientist Theodor Boveri in 1914 that laid ground for later identification of abnormalities in genomic DNA as the molecular basis for cancer (Boveri 2008). According to Boveri's hypothesis, chromosomal abnormalities caused the transition from normal to malignant proliferation. The identification of DNA as the transforming principle (Avery et al. 1944) and the elucidation of structure of the double helix (Watson & Crick 1953) paved way for the study of DNA in cancers. Other significant advances in 1950s, which included the cytogenetic characterization of mammalian chromosomes and the ability to culture cells *in vitro* (Moorhead et al. 1960), enabled the identification of correct

chromosomal number in humans (2n=46) (Tjio & Levan 2010). The first recurrent chromosomal aberration, identified in seven patients with chronic myeloid leukemia (CML) using cancer cytogenetics, was termed the Philadelphia (Ph) chromosome. It was a small karyotypic marker that replaced one of the four smallest autosomes (Nowell & Hungerford 1960). This discovery validated the somatic mutation theory proposed by Boveri half a century ago and generated great interest for subsequent characterization of cancer chromosomes.

1.1.1. Chromosomal banding methods

The next major advance in studying cancer cells came with the advent of chromosomal banding techniques, developed during the early 1970s. It was observed that individual chromosomes presented with distinguishable banding patters, when isolated during the highly condensed metaphase stage of the cell cycle and stained with a fluorophore (Caspersson et al. 1968; Caspersson et al. 1970). This observation led to the development of a series of banding methods, each using a different dye or other pre-treatment of DNA, which were used to study metaphase chromosomes from normal and tumor cells amenable to culturing (Manolov & Manolova 1972; De la Chapelle et al. 1972; Rowley 1973a; Francke 1976; Muleris et al. 1985). Subsequently, the Ph chromosome was identified to result from a translocation between chr9 and chr22 (Rowley 1973b). Banding analysis enabled the detection of cancer-specific chromosomal aberrations in a wide variety of hematological malignancies (Manolov & Manolova 1972; De la Chapelle et al. 1972), and with some limitations, in solid tumors as well (Kovacs 1978; Limon et al. 1986). Typically, 400 bands are detected using metaphase chromosomes, which limit the resolution of detected aberrations to approximately 10 Mb in size (Shaffer & Bejjani 2004). Still, banding methods allow single-cell, pan-genome analysis and

are commonly used in cytogenetic investigations. The development of recombinant DNA technologies and molecular cloning approaches in 1970s and 1980s allowed direct manipulation of DNA molecules (Weiss & Richardson 1967; Smith & Wilcox 1970; Kelly & Smith 1970; Jackson et al. 1972; Cohen et al. 1973; Lobban & Kaiser 1973; Nathans & Smith 1975). Using these advances, it was identified that the translocation in Philadelphia chromosome resulted in a fusion between two genes, ABL1 on chr9 and BCR on chr22 (de Klein et al. 1982; Groffen et al. 1984; Heisterkamp et al. 1985). The Ph chromosome story is a wonderful scientific story of targeted therapy in cancer. The development of Imatinib, a tyrosine kinase inhibitor, to target the BCR-ABL fusion in CML is the first example of targeted therapy against a specific fusion gene in cancer (Druker 2004). Around the same time, the discovery of first proto-oncogene (Stehelin et al. 1976) and first cancer specific point mutation in the gene HRAS (Reddy et al. 1982) established genetic aberrations as the cause of malignant transformation.

1.1.2. Fluorescence *in situ* hybridization (FISH) methods

The development of *in situ* hybridization methods using radio-labeled probes in late 1960s (Gall & Pardue 1969; Pardue & Gall 1969), which led to later development of FISH methods (Bauman et al. 1980), opened new frontiers for cancer genome analysis (Trask 2002). In FISH based methods, metaphase or interphase chromosomal spreads are denatured, and single-stranded DNA probes with fluorescent tags are then allowed to hybridize with their complementary genomic sequences on these chromosomes (Lichter et al. 1988). While metaphase FISH requires viable cells from which metaphase chromosomal spreads can be prepared during cell growth, interphase FISH can be performed with post-mitotic cells or preserved cellular material (Vorsanova et al. 2010). In the simplest version, DNA probes can

detect copy number of the targeted sequence. However, more than one probe with different fluorescent labels can be used in the same experiment. Such multiplexing allows probing of many genomic loci at the same time (Hopman et al. 1986). Additionally, colocalization of different probes can serve as an indicator of a fusion or other chromosomal rearrangement. The ability to design probes for entire genomes led to the development of multicolor karyotyping, which is epitomized by spectral karyotyping (SKY) (Schröck et al. 1996). In SKY karyotyping, a large number of differentially labeled probes, which target all chromosomes in the genome, allow chromosomal identification by assigning each pair of homologous chromosomes a unique spectral signature. When the probes are hybridized to the genome, they paint each chromosome with a different color, which is used to determine structural rearrangements like translocations etc.

1.1.3. Comparative genome hybridization (CGH) methods

Developed in early 1990s, CGH enabled genome-wide analysis of copy number gains and losses in tumor cells. For CGH, DNA is first extracted from a test and a normal reference sample. These DNA preparations are then labeled with different fluorophores, mixed and allowed to hybridize to complementary probe sequences. By image analysis and comparison of relative fluorescence intensities between both fluorophores at each probe location, deleted or amplified genomic regions in the test genome can be identified (Kallioniemi et al. 1992). In the beginning, the probe sequences were metaphase chromosomal spreads (hence chromosomal CGH). Later, probe DNA molecules were spotted onto a glass slide surface (hence called arrayCGH). Array CGH probes could be genomic DNA molecules cloned in bacterial artificial chromosomes (BAC arrays) or synthesized oligonucleotides (oligonucleotide arrays; Solinas-

Toldo et al. 1997; Pinkel et al. 1998; Bignell et al. 2004; Pinkel & Albertson 2005). Although limited in resolution and their ability to identify balanced structural variation, CGH methods are amenable to automation and fairly easy to use. Consequently, they became commonly used in cancer cytogenetics (Albertson 2003; Schwaenen et al. 2004).

1.1.4. DNA sequencing based methods

With the technological advances that enabled quick and cost-effective analysis of whole human genomes (Margulies et al. 2005; Turcatti et al. 2008; Bentley et al. 2008; Metzker 2010; Rothberg et al. 2011) since completion of the Human Genome Project (Lander et al. 2001; Venter et al. 2001; International Human Genome Sequencing Consortium 2004), it became possible to characterize the gamut of variation in normal (1000 Genomes Project Consortium 2010; Abecasis et al. 2012) and cancer genomes (Weinstein et al. 2013). These technological advances led to the development of many next-generation sequencing systems, which eliminated costly and time consuming process of generating clone based libraries for whole genomes. Instead, whole genome libraries could be prepared in single tubes simply by fragmenting genomic DNA and modifying resulting fragments to prepare them for sequencing, thereby saving both cost and time. Additionally, the human reference genome provided a template to which sequencing reads could be compared and scored for variants, thereby eliminating the need for a finished *de novo* assembly for every genome. Consequently, we have witnessed substantial improvement in our understanding and appreciation of genome instability in different kinds of cancers.

Initially, DNA sequencing based cancer genomics efforts were focused at re-sequencing a subset of or the entire exome using PCR based amplification followed by Sanger sequencing

(Wood et al. 2007; Greenman et al. 2007; TCGA 2008; Ding et al. 2008). The Cancer Genome Atlas's (Hartwell & Lander 2005) pilot project analyzed 601 genes for somatic variation across 91 glioblastoma samples and identified significant variants in previously known (ERRB2, TP53) and unknown (PIK3R1) genes (TCGA 2008). Ding *et al.* studied 623 genes in lung cancer patient samples to identify 26 significantly mutated genes, which affected key signaling pathways (Ding et al. 2008). In 2008, whole genome sequencing was first employed to study a cancer genome, an acute myeloid leukemia (AML-M1) genome with normal cytogenetic status. In this genome, Ley and colleagues identified 10 nonsynonymous somatic single nucleotide variants, 2 of which were previously known to play a role in AML pathogenesis (Ley et al. 2008). This study established the feasibility of studying whole tumor genomes using next generation DNA sequencing. Around that time, a number of other studies reported sequencing of other cancer genomes and presented an initial glimpse into the mutational landscape of these cancers (Shah et al. 2009; Mardis et al. 2009; Pleasance, Stephens, et al. 2010; Pleasance, Cheetham, et al. 2010; Lee et al. 2010). With further increases in throughput, decrease in cost and computational ease of analysis, multi-center projects for sequencing of cohorts of patients for many tumor types, spearheaded mostly by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), were successfully completed (TCGA 2008; Chapman MA et al. 2011; TCGA 2011; TCGA 2012b; TCGA 2012a; TCGA 2012c; Kandoth, Schultz, et al. 2013; TCGA 2013; Kandoth, McLellan, et al. 2013; TCGA 2014). These studies have revealed widespread inter-tumor and intra-tumor heterogeneity of the mutational landscape, which targets common downstream pathways. We now know that the average number of significantly mutated genes varies across cancers (Kandoth, McLellan, et al. 2013). The most

complex colorectal and lung cancers harbor hundreds to thousands of somatic mutations per cancer genome while the least complex leukemias and pediatric cancers carry ~10 somatic mutations per cancer genome (Lee et al. 2010; Vogelstein et al. 2013). As expected, environmental factors were confirmed to play an important role in genome variation, as was illustrated by the finding that smokers with lung cancer had 10 times as many mutations as nonsmokers with lung cancer (Govindan et al. 2012).

Recent reports have also introduced novel patterns of genome variation in cancers. Chromothripsis, where a single event of genome shattering and reassembly results in oscillating copy number changes that localize to one or a few chromosomes, has been found to occur in 2-3% of all cancers and more frequently in bone cancers (Stephens et al. 2011). Kataegis, where a cluster of somatic single nucleotide variants colocalize with somatic structural variants, was first reported for breast cancers (Nik-Zainal et al. 2012) and then for pancreatic cancer and lung adenocarcinoma as well (Alexandrov et al. 2013). Beyond genomic, transcriptomic (Wu et al. 2008; Maher et al. 2009) and epigenetic (Jones & Baylin 2007) mechanisms have also been implicated in cancer pathogenesis, but a rigorous discussion of these mechanisms is beyond the scope of our current work.

These technological advances have also enabled the genomic analysis of multiple time-points during tumor evolution in patients, thereby shedding some light on the mutational processes that accompany tumor progression. We now know that tumors evolve through an ongoing process of genomic instability and clonal selection, leading to clonal expansion of the most favorable clone (Stratton et al. 2009; Stratton 2011). Tumors can evolve linearly, accumulating mutations as they progress (Shah et al. 2009). Alternatively, they can evolve in a

branching fashion, where a minor sub-clone carrying most of primary tumor mutations survives, gains additional mutations and expands later at relapse (Ding et al. 2012). These mechanisms are reminiscent of a Darwinian model of evolution, which supports the growth of the fittest clone, and can explain genomic mechanisms underlying drug response, drug resistance and metastasis. For instance, in colorectal carcinoma, a small number of mutations transformed a highly invasive cancer cell into a metastatic one (Jones et al. 2008). In uveal melanoma, a single mutational event in BAP1 (BRCA1-associated protein 1) distinguished metastatic tumor from a non-metastatic one (Harbour et al. 2010). Mutations connected to chemotherapeutic resistance have been reported in glioblastomas where inactivating mutations or loss of the mismatch repair protein MSH6 leads to resistance towards alkylator based chemotherapy (Hunter et al. 2006; Cahill et al. 2007).

While we are still far away from the initially proposed goal of curing cancers with a better understanding of cancer genomes, the knowledge gleaned from current analysis of cancer genomes has substantiated the potential of whole genome sequencing for diagnostic, prognostic and therapeutic decision-making in cancer. In non-small-cell lung cancer, the T790M mutation in EGFR leads to relapse during therapy with tyrosine kinase inhibitors gefitinib or erlotinib (Kobayashi et al. 2005; Pao et al. 2005). On the other hand, in lung adenocarcinomas, activating mutations and amplifications of EGFR lead to significant survival benefit upon treatment with gefitinib and erlotinib. An inverse effect has also been observed in wild-type EFGR patients (Mitsudomi et al. 2010). Consequently, EFGR is routinely sequenced in lung cancer patients to predict the efficacy of tyrosine kinase inhibitors before starting treatment (Wong et al. 2011). In lung cancers with EML4-ALK rearrangement (Soda et al. 2007), ALK

inhibitors, particularly crizotinib, significantly influence tumor response (Kwak et al. 2010). Colorectal cancer patients with KRAS mutations have low response to anti-EFGR therapies, while the ones with no mutations along with KRAS amplifications generally have a much higher response (Benvenuti et al. 2007). In chronic myeloid leukemia (CML), imatinib and the next-generation of small molecule inhibitors are commonly used to target the constitutively active ABL kinase resulting from the canonical BCR-ABL t(9;22) translocation (Druker 2008). In breast cancer patients with HER2 amplification, trastuzumab leads to good tumor response (Esteva et al. 2010). Specific inhibitors have been developed to target the V600E mutation in BRAF, which leads to constitutive activation of BRAF kinase in many cancers (Chapman PB et al. 2011). Overall, all of these examples highlight how genomic information gleaned from cancer genomes is being utilized to make clinical decisions in cancer treatment. With an improved understanding of molecular processes underlying cancer pathogenesis, we will be able to design better treatments in near future.

1.2. Structural variation in normal and cancer genomes

The previous section summarized the methods used for and the information gained from studying genomic abnormalities in cancer genomes. Using these methods, primarily hybridization and sequencing based techniques, recent work has identified widespread copy number and structural changes in both, normal and cancer genomes. In the following section, I have reviewed recent work that has been done to identify such changes in human genomes. With the growing understanding that cancer genomes harbor widespread structural variation and rearrangements, I have highlighted the need to routinely analyze cancer genomes for structural variation.

1.2.1. Structural polymorphisms in normal genomes

Beginning in 2004, a number of studies identified widespread structural polymorphisms in normal human genomes. Using a technique called representational oligonucleotide microarray analysis (ROMA), Sebat and colleagues identified 221 copy number changes greater than 100 kb in size representing 76 unique loci in 20 individuals (Sebat et al. 2004). ROMA, analogously to comparative hybridization methods, measures the relative concentration of DNA in two samples by hybridizing differentially labeled samples to a set of probes. Later, Iafrate *et al.* used BAC array based comparative genomic hybridization to identify 255 large scale copy number variations (copy number incongruent clones) in 55 unrelated individuals, with 102 (41%) changes observed in more than one individual, 24 changes observed in more than 10% individuals and on average, 12.4 large scale copy number variants per individual (Iafrate et al. 2004). Using SNP genotyping, which improved the resolution of detection, Conrad and colleagues detected 586 deletion polymorphisms ranging in size from 300 bp to 1200 kb in parent-offspring trios (Conrad et al. 2006). Although limited in resolution, these seminal studies highlighted the presence of rare and common polymorphic structural differences in normal human genomes.

More recently, DNA sequencing based methods, particularly paired-end sequencing have been used for the identification and characterization of basepair level resolved structural changes in human genomes. In paired-end sequencing, a library of DNA fragments, closely distributed in size, is prepared. For instance, in a fosmid based library, the average fragment size lies around 40 kb. In more recent DNA preparations, average fragment sizes of 250 bp to 10 kb have been used. Short stretches of these fragments (~100 bp) are sequenced from the ends

to generate paired-end data. The paired end data, which contains information about the sequence at the ends and the physical distance separating these sequences, can then be used in many ways to computationally identify structural variants (reviewed in Alkan et al. 2011). Discordant read pairs identified using read-pair approaches bear indicators of deletions, inversions and translocations etc. based on incongruent insert size and/or orientation (Chen et al. 2009). Copy number differences identified using read-depth based approaches can identify amplifications and deletions (Abyzov et al. 2011). Split alignments from split-reads can be used to identify deletions, insertions, inversions and translocations etc. (Ye et al. 2009). Finally, *de novo* assembly based approaches can be used to assemble genomic rearrangements (Simpson et al. 2009; Luo et al. 2012).

In one of the earlier studies, split capillary reads of DNA resequencing traces from three diverse human populations were used to identify more than 400,000 indels, ranging from 1 bp to 9989 bp in size (Mills et al. 2006). Eichler and colleagues used paired-end sequencing with fosmid libraries to identify 297 structural variants greater than 8 kb in size in one human genome. These variants included 102 deletions, 139 insertions and 56 inversion breakpoints. The variant regions were found to be enriched in or near repetitive DNA, with 55% (163 of 297) of the identified variants mapping to segmentally duplicated regions that otherwise represent only 5.3% of the genome (Tuzun et al. 2005). Later, Korbel *et al.* used high throughput and massive paired-end mapping (PEM) of 3 kb fragments to identify 1297 structural variants, ranging in size from 3 kb to megabases, in two human genomes (Korbel et al. 2007). These variants included 853 deletions, 322 insertions and 122 inversions; 45% of the identified variants were shared between both genomes. Consistent with previous (Tuzun et al. 2005) and

later (Sudmant et al. 2010) work, structural variants were found to be strongly associated with segmental duplications and L1 LINE elements. Previous work from our lab also revealed thousands of structural variants in a complete hydatidiform mole and three lymphoblast-derived cell lines using Optical Mapping (Teague et al. 2010).

Recent large scale studies have further expanded our understanding of copy number and structural changes in human genomes (Kidd et al. 2008; Conrad et al. 2010; Sudmant et al. 2010; Mills et al. 2011). Using tiling oligonucleotide microarrays with 42 million probes, a recent study identified 11,700 copy number variations (CNVs) greater than 443 basepairs across 40 human samples (Conrad et al. 2010). Using short read sequencing data, Sudmant and colleagues identified large copy number changes in many genes across 159 human genomes. They identified 56 most variable gene families with a median copy number of 5-368 across these human samples. Not surprisingly, majority of the events greater than 50 kb in size (55%, 522/952) overlapped segmental duplications (Sudmant et al. 2010). Phase I of 1000 genomes project analyzed sequencing data from 185 human genomes and generated >28,000 variants greater than 50 bp in size, with more than 53% variants identified at nucleotide resolution (Mills et al. 2011). Following up on these large-scale studies, genomic aberrations larger than 50 bp in size are now defined as structural variants.

Generally speaking, copy number and structural polymorphisms are now considered a normal component of human genome structure. Such differences have been implicated in disease susceptibility and thought to have an effect on complex human traits. It is largely accepted that these polymorphic differences affect a larger number of base pairs than commonly analyzed single nucleotide polymorphisms (SNPs) and small indels (Korbel et al.

2007; Zhang et al. 2009; Conrad et al. 2010). However, these polymorphisms are known to be selectively enriched in segmentally duplicated and other low complexity regions of the genome (Tuzun et al. 2005; Korbel et al. 2007; Sudmant et al. 2010), which complicates their analysis using current sequencing based approaches, as has been discussed later in this chapter.

1.2.2. Structural variation and genome rearrangement in cancer genomes

Using one or more of paired-end sequencing based structural variation identification approaches (Alkan et al. 2011), somatic structural rearrangements have been identified in lung (Campbell et al. 2008), breast (Stephens et al. 2009; Ding et al. 2010), pancreatic (Campbell et al. 2010) and other cancers. Recent work from our lab used Optical Mapping to discover and characterize a large number of structural variants in oligodendroglioma (Ray et al. 2013). Apart from cancers, structural and copy number aberrations are commonly known to be associated with autism spectrum disorders (Sebat et al. 2007), rare neurodevelopmental diseases (Girirajan et al. 2011), psoriasis (Hollox et al. 2008) and other complex human disorders (Zhang et al. 2009). These structural variants arise because of homology-mediated and non-homology mediated mechanisms (Hastings et al. 2009), and affect a wide variety of genes with functional implications in telomere stability, cell cycle control and other cellular processes (Campbell et al. 2010).

While the advances in genomic analysis methods mentioned above have enabled and the aforementioned focal studies have highlighted the presence of widespread structural variation in cancer genomes, routine analysis of structural variation alongside single nucleotide and copy number variation precludes us because of many limitations that are associated with the use of these methods. Classical cytogenetic methods have very low resolution and

sensitivity. For highly complex malignancies, data analysis and interpretation becomes a problem. Karyotyping also requires actively dividing cells, which is not achievable for many solid and hematological cancers, thereby limiting analysis. Although FISH provides greater sensitivity and resolution, it requires pre-defined probes, which limits discovery. Additionally, only a few loci can be evaluated in one set of experiments. Although multiple loci can be investigated at the same time, it increases the complexity of testing. Microarray based genome-wide analysis methods are limited in resolution (~25 kb for highest resolution arrays) and in their ability to identify and characterize balanced events. They also lack the ability to provide structural and contextual information at copy number breakpoints.

While DNA sequencing based methods conceptually promise comprehensive genomic variation analysis, they fail to do so in practice. A major reason for this incongruence is that fact that structural variants, because of the mechanisms from which they occur, tend to lie in genomic regions with repetitive elements such as segmental duplications and transposons (Sudmant et al. 2010). Because of the inability of short-read DNA sequencing data to uniquely differentiate these regions, true positives are difficult to discern in these regions. Additionally, false negative rates as high as 37% have been reported (Biankin et al. 2012), which could still be an underestimate. Also, most of the DNA sequencing based structural variation calling algorithms show high false positive rates (Abel & Duncavage 2013), which makes it difficult to distinguish real genomic aberrations from computational artifacts. It is because of these reasons that different sequencing based structural variation calling algorithms show little overlap (Alkan et al. 2011). Also, there is generally a bias towards detecting deletions over insertions because inserted sequences are more difficult to identify using these reference

sequence based approaches. Finally, validating and ascertaining functional consequences of structural variation remain a challenge. These variants routinely encompass multiple genes, and identifying which ones are functionally relevant requires time-consuming follow-up experimentation. In chapter 2 of this thesis, I will describe the application of Optical Mapping to study structural variation in multiple myeloma. Through our analysis, I will discuss how Optical Mapping, because of its unique approach to genome analysis, is able to overcome most of the limitations with currently used methods.

1.3. Optical Mapping for genome analysis

Optical Mapping is a high-throughput, single-molecule system that generates ordered restriction maps (Rmaps) from high molecular weight genomic DNA molecules, ranging in size from 300 kilobases to megabases. These Rmaps are then used for the construction of genome-wide physical restriction maps using computational approaches, which provides insights into long range genome structure and genome variation. The mapping system uses unamplified genomic DNA as substrate, thereby yielding a view of the genome free of cloning or amplification biases (Schwartz et al. 1993; Meng et al. 1995; Cai et al. 1995; Anantharaman et al. 1997; Cai et al. 1998; Jing et al. 1998; Anantharaman et al. 1999; Skiadas et al. 1999; Lai et al. 1999; Lin et al. 1999; Lim et al. 2001; Perna et al. 2001; Zhou et al. 2003; Zhou et al. 2004; Dimalanta et al. 2004; Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Zhou et al. 2007; Kidd et al. 2008; Zhou et al. 2009; Antonacci et al. 2010; Teague et al. 2010; Chen et al. 2012; Ray et al. 2013).

An outline of Optical Mapping is provided in figure 1-1. The first step in Optical Mapping is DNA extraction. Because high molecular weight DNA is required as a substrate for Optical

Mapping, very gentle DNA extraction methods like liquid lysis or DNA inserts (Schwartz & Cantor 1984) are commonly used. Next, DNA is presented on positively-charged surfaces (Meng et al. 1995; Cai et al. 1995; Hu 1997; Lim et al. 2001) *via* capillary flow in microchannels (Dimalanta et al. 2004). DNA presentation encompasses two goals: elongation and immobilization. DNA elongation allows the imaging of molecular cleavage events once intact DNA molecules are digested using restriction endonucleases, and is an important requirement for generation of Rmaps. DNA immobilization serves to fix DNA in place, which is important to ensure that *i*) the linear order of DNA fragments from each DNA molecule is preserved; *ii*) the digested molecules can be imaged easily; and *iii*) the fragments generated after restriction digestion do not desorb and get lost before imaging. Both these steps, elongation and immobilization, are carefully controlled to ensure that the biochemical action of restriction endonucleases is preserved and that the DNA molecules are optimally stretched out to be able to generate useful data. Following presentation, the DNA molecules are digested with a restriction endonuclease of choice, stained using an intercalating fluorochrome (Cai et al. 1995) and then imaged using automated laser-illuminated epifluorescence microscopy systems (Jing et al. 1998; Skiadas et al. 1999; Zhou et al. 2004; Dimalanta et al. 2004). Upon digestion, the cut sites present as gaps that are formed between fragments due to DNA relaxation at cut ends (Schwartz et al. 1993; Meng et al. 1995). Custom-developed software for image analysis and machine vision is then used to obtain ordered restriction maps from single DNA molecules (Dimalanta et al. 2004; Teague et al. 2010).

Once a large dataset of Rmaps has been collected using the Optical Mapping system, a computational pipeline that uses Bayesian inference approaches (Anantharaman et al. 1999)

and cluster computing is used to assemble the Rmaps into genome-wide contigs and generate genome-wide consensus maps (Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Teague et al. 2010). Finally, the consensus maps are used in multiple ways to learn about genome structure. For instance, Optical Mapping has been used to aid/guide DNA sequencing based genome assembly projects (Lai et al. 1999; Lin et al. 1999; Lim et al. 2001; Armbrust et al. 2004; Zhou et al. 2007; Zhou et al. 2009; Wei et al. 2009), for bacterial comparative genomics (Zhou et al. 2004) and for the analysis of structural variation in normal (Teague et al. 2010) and cancerous (Ray et al. 2013) human genomes.

1.3.1. Optical Mapping - Development

Optical Mapping has been made possible by parallel development and seamless integration of a number of experimental and computational components, the current state of which has been described above. In the following sections, I will detail the development of these components.

1.3.1.1. DNA immobilization

For the first version of Optical Mapping, Schwartz and colleagues (Schwartz et al. 1993) used an agarose gel based method for DNA immobilization. Fluid flow was used to stretch out DNA molecules dissolved in molten agarose. The subsequent gelation process fixed DNA molecules in place and prevented relaxation of DNA molecules to a coiled coil conformation. Finally, the agarose gel matrix is sufficiently porous to restriction endonucleases and their cofactors, thereby allowing enzymatic activity in gel-trapped and elongated DNA molecules. An elongation of 20 - 30% of contour length was achieved, which potentiated automatic ordering of the fragments in individual DNA molecules. The authors tracked restriction digestion with

time-lapsed imaging and noted that cut sites presented themselves as growing visible gaps in elongated DNA. Also, bright and condensed pools of DNA formed on the ends of fragments flanking the cut site, due to the local relaxation of DNA molecules. With subsequent data analysis, ordered restriction maps of *S. cerevisiae* chromosomes were constructed, and found to correspond well with similar maps generated in prior studies (Link & Olson 1991) or those generated using pulsed field gel electrophoresis (Schwartz & Cantor 1984). However, this mode of presentation limited throughput and efficiency because of the cumbersome time-lapse imaging. It also suffered from low resolution (~20 kb) because of relaxation and subsequent disappearance of smaller DNA fragments.

An improved presentation scheme used positively-charged polylysine-treated glass surfaces for DNA immobilization (Meng et al. 1995). The DNA molecules were fixed on derivatized glass surfaces by sandwiching the DNA sample (which generated flow) between the surface and a glass slide. The polylysine molecular weight and its concentration on glass surfaces were optimized to achieve DNA fixation, elongation and accessibility to enzymatic action, all without the use of chemical crosslinking. Instead, DNA presentation relied on electrostatic interactions between the negatively charged DNA chains and positively charged polylysine modified surfaces. A small hole in the slide was used to provide the restriction enzyme and magnesium ions to start the digestion process. Because DNA molecules were immobilized onto surfaces and remained in focus, even after restriction digestion, it obviated the need for time-lapsed imaging. This improvement made the process much more convenient, increased size resolution to 800 bp and reflected an important advancement towards the goal of high throughput, fully automated Optical Mapping.

Over the years, different positively-charged silanes have been used for surface derivatization and DNA immobilization. APTES (3-aminopropyltriethoxysilane) (Cai et al. 1995), APDEMS (3-aminopropyldiethoxymethylsilane) (Lim et al. 2001), trimethyl silane (N-trimethylsilylpropyl-N,N,N-trimethylammonium chloride) and vinyl silane (vinyltrimethoxysilane) (Hu 1997) have since been used. Currently, a combination of trimethyl and vinyl silanes is routinely used to immobilize DNA for Optical Mapping.

1.3.1.2. DNA elongation

DNA elongation has generally been accomplished by enabling flow of a DNA sample solution. Significant progress has been made to the flow processes which enable DNA elongation on positively charges surfaces. The sandwich mount has been used in two forms: *i*) By squeezing the DNA sample between a surface and a glass slide (Meng et al. 1995; Cai et al. 1998); and *ii*) By adding the DNA sample to the edge between the surface and a glass slide (Jing et al. 1999). Upon addition of sample solution to the edge, the solution spreads into the space between the surface and the slide by capillary action, promoting the elongation of many DNA molecules in same direction, thereby reducing inter-molecule overlaps and noise. A new approach, termed fluid fixation, where fluid flow within tiny, evaporating droplets was used to elongate and fix DNA to charged surfaces, was used for successful mapping of bacteriophage lambda DNA (Jing et al. 1998) and long-range PCR products from a human gene (Skiadas et al. 1999). This approach was the first automated system for analysis of single DNA molecules, which comprised spotting engines, image acquisition systems, machine vision and assembly algorithms that collectively, greatly boosted throughput. Additionally, rehydration of the spotted DNA molecules with restriction enzyme buffer enabled restriction endonuclease action.

Finally, development of a microfluidic device allowed for massively-parallel, high-throughput deposition of single DNA molecules on derivatized glass surfaces that enabled the physical mapping of many bacterial, fungal, plant and mammalian genomes (Dimalanta et al. 2004). DNA presentation was achieved by capillary flow of DNA solution in microchannels, which are formed at the interface of a glass surface and a microfluidic device made from polydimethylsiloxane (PDMS) using soft lithography approaches (McDonald et al. 2000; Whitesides et al. 2001; Ng et al. 2002). During capillary flow, DNA elongates and immobilizes on the glass surface *via* electrostatic interactions and flow effects while staying biochemically accessible.

1.3.1.3. Restriction endonuclease digestion

Currently, restriction digestion is accomplished by addition of an enzyme solution on top of glass surfaces with immobilized DNA, and incubation at an optimum temperature in a humidified chamber. It was observed that after enzymatic action, smaller fragments would desorb from the surface, possibly because of lower cumulative electrostatic interactions on account of their small size. To ameliorate this desorption, a polyacryamide overlay was used atop presented DNA (Hu 1997; Zhou, Deng, et al. 2002). The polyacrylamide overlay allows the use of longer restriction enzyme incubation times in order to achieve optimal digestion yields without over-relaxation of cut ends or loss of small fragments.

1.3.1.4. DNA staining

Initially, fluorochromes like DAPI (4,6-diamino2-phenylindole dihydrochloride) (Schwartz et al. 1993) or ethidium homodimer (Meng et al. 1995) were used for DNA staining. However, it was later found that YOYO-1, an oxazole yellow homodimer generated images that were much

brighter and better in contrast (Cai et al. 1995). This shortened imaging time and also provided better images for data analysis. Also, high salt conditions, which severely diminished fluorescence emission of ethidium stained molecules, had no effect on the luminosity of YOYO-1 stained DNA molecules, even under severe fixation conditions. The improvement in image quality also improved sizing results, particularly for smaller fragments. Addition of 2-mercaptoethanol prevented photo-damage of surface presented molecules. Currently, YOYO-1 mixed with 2-mercaptoethanol is used to stain DNA molecules.

1.3.1.5. Imaging and image processing

During the early days of Optical Mapping, manual imaging and image processing were used to generate useful map data (Schwartz et al. 1993; Meng et al. 1995). Either fluorescence intensity or fragment lengths were used to determine fragment sizes. However, as the scale and scope of Optical Mapping has increased from mapping yeast chromosomes (Schwartz et al. 1993) to microbial (Lin et al. 1999), plant (Zhou et al. 2007; Zhou et al. 2009; Wei et al. 2009) and whole human genomes (Teague et al. 2010; Ray et al. 2013), concomitant improvements in imaging and image processing have been accomplished (Skiadas et al. 1999; Dimalanta et al. 2004). Optical Map Maker (Jing et al. 1998) was first such automated system and integrated all of the workstation functions, which included movement of the microscope stage, focus and image collection. It was integrated with a storage system that allowed later image processing and controlled access to all collection and processing activities and data. Different automated systems were developed for image collection and processing during the development of Optical Mapping. The image processing software has evolved from a completely manual (Visionade), to

a semi-automatic (semi-Autovis; Lim et al. 2001) to a fully automatic process (Pathfinder and others) (Zhou et al. 2004; Dimalanta et al. 2004).

In current state, DNA molecules are imaged on a Zeiss 135 M inverted microscope (Carl Zeiss Microimaging), equipped with a 63x Zeiss oil immersion objective. Illumination is achieved *via* 488 nm emission line from an argon-ion laser (Spectra-Physics). The microscope is also fitted with a high speed CCD camera and shutter, stage and focus controllers (LUDL Electronics Products). Custom in-house software allows automated imaging of an entire array of microchannels with very little setup time. Once the images have been collected, they are automatically processed with two pieces of software. The first one uses a bright image to flatten raw micrographs, which accounts for uneven illumination, and then overlaps and merges the images collected from each microchannel into a single composite image, which can then be used to generate useful data from DNA molecules scanning multiple frames. The second software, called Pathfinder, identifies single molecules of DNA and then individual restriction fragments. It sizes the fragments based on relative fluorescence intensities, relative to a sizing standard (generally a lambda genome for which the digestion pattern is known), which is mixed with DNA sample and presented on glass surfaces. This process generates ordered restriction maps (Rmaps), which are then used to create physical restriction maps for whole genomes.

1.3.1.6. Generation of physical restriction maps from Rmaps

Single molecule Rmaps have errors of their own: missing cut sites due to incomplete digestion, extra cut sites due to non-specific digestion, fragment sizing errors because of experimental artifacts and missing fragments which desorb and disappear because of their

small size and low electrostatic interactions (discussed in Teague et al. 2010). The errors in single molecule maps are minimized by collecting many map observations from each genomic locus and generating consensus maps, which reflect the genome structure of the sample being analyzed (Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Teague et al. 2010).

From the very beginning, many molecules were collected, overlaid and then the information was averaged to generate consensus maps. When the genomes analyzed were smaller, for instance with yeast chromosomes (Schwartz et al. 1993), bacteriophage clones (Meng et al. 1995) and yeast artificial chromosomes (Cai et al. 1995), this process was amenable to manual analysis. However, as collection of much larger amounts of data became possible for larger genomes because of experimental advances, it necessitated the development of automated computational methods for generating physical maps. The development of Gentig (Anantharaman et al. 1999) was an important step in automated assembly of Rmaps. Gentig uses an approximation algorithm to reliably overlap maps derived from single DNA molecules. The overlapped maps can then be assembled into contigs using a heuristic Bayesian algorithm for finding the best scoring set of contigs. The first whole genome maps assembled using Optical Mapping of *Deinococcus radiodurans* were initially generated using manual assembly and later using Gentig based assembly; the manual and Gentig based assembly approaches were shown to be highly congruent (Lin et al. 1999). For smaller bacterial and fungal genomes, simple computational needs meant single computers could be used in the process. However, for much larger plant and mammalian genomes (Zhou et al. 2007; Zhou et al. 2009; Wei et al. 2009), it required the development of cluster computing approaches to address

the computational needs of assembling many thousands to millions of Rmaps (Zhou et al. 2007). *De novo* assembly approaches have been previously used in projects like rice (Zhou et al. 2007) and maize (Zhou et al. 2009) genome mapping where a good reference sequence was not available. For Optical Mapping of human genomes, where a high quality reference sequence is available, a reference-based iterative assembler was developed for the assembly of Rmap datasets (Valouev, Schwartz, et al. 2006; Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Teague et al. 2010).

During first iteration of iterative assembly, single molecule Rmaps are clustered by pairwise alignment to an *in silico* restriction map derived from the reference sequence. The clusters are then locally assembled using a maximum-likelihood Bayesian assembler (Anantharaman et al. 1999) to generate new hypothesis consensus maps, which are used as reference maps for subsequent rounds of iterative assembly. As iterations proceed, the hypothesis consensus maps are updated and look more representative of the genome being analyzed. Fig. 1-2 illustrates the computational pipeline used for iterative assembly. Generally, eight iterations are used to generate an accurate and comprehensive human genome assembly (described in Teague et al. 2010).

1.3.2. Applications of Optical Mapping

Optical map assemblies provide long range structural information about the genome. Consequently, they generate a scaffold that can be used to verify or guide DNA sequencing based genome assemblies. Optical Mapping was first used to verify sequencing based chromosomal (Jing et al. 1999) and genome assemblies (Lin et al. 1999). With an increase in throughput, it was used to generate physical assemblies to aid sequencing based genome

assembly for many microbial genomes. These include some bacterial genomes like *Deinococcus radiodurans* (Lin et al. 1999), *Escherichia coli* O157:H7 (Lim et al. 2001), *Yersinia pestis* (Zhou, Deng, et al. 2002) and *Rhodobacter sphaeroides* 2.4.1 (Zhou et al. 2003). Physical assemblies have also been generated for the protozoan *Plasmodium falciparum* genome (Lai et al. 1999) and algal *Thalassiosira pseudonana* genome (Armbrust et al. 2004). By comparing different bacterial strains to identify genomic differences, Optical Mapping was used for comparative genomics (Zhou et al. 2004). More recently, plant genomes like rice (Zhou et al. 2007) and maize (Zhou et al. 2009; Wei et al. 2009) and normal (Teague et al. 2010) and cancerous (Ray et al. 2013) human genomes have been mapped. These assemblies have helped in validation of sequencing based assemblies and have also provided high-resolution scaffolds for gap closure and for correcting sequencing based assembly errors (Reslewic et al. 2005). Optical Mapping has also been used to study genome-wide methylation patterns (Ananiev et al. 2008) and single molecule analysis of transcription (Wu & Schwartz 2007).

Optical Mapping of human genomes has uncovered a wide array of structural variation in these genomes. Teague et al. identified thousands of structural variants, ranging in size from a few kilobases to megabases in a complete hydatidiform mole and three lymphoblast-derived cell lines (Teague et al. 2010). The authors also identified many structural variants that were inaccessible to other genomic analysis platforms. Later, Ray et al. studied tumor genomes from two oligodendroglioma patient samples, the first use of Optical Mapping to study a solid tumor genome, to reveal many somatic structural variants and copy number heterogeneity (Ray et al. 2013).

1.3.3. Nanocoding

Although very useful in discerning genome structure, Optical Mapping is limited in its identification of structural variants mostly to events larger than 3 kb in size. There are two major factors which result in this limitation: *i*) Many fragments smaller than 3 kb desorb from surfaces due to lower cumulative electrostatic binding and are consequently, not assembled during assembly; and *ii*) Rmaps present with some sizing errors, which are more pronounced for smaller fragments because of experimental issues. Also, these factors require that larger raw datasets be collected, which means longer turnaround time for sample analysis. To address these issues, our laboratory has developed an improved genome analysis system called Nanocoding (Jo et al. 2007; Jo et al. 2009), which promises to improve the resolution of structural variant detection to 1 kb, and significantly reduce turnaround times (unpublished results).

Nanocoding uses fluorochrome incorporation at single-stranded sequence-specific nick sites to generate single molecule maps called Nmaps. Briefly, long genomic DNA is healed to remove existing non-specific nicks, nicked using a single strand nicking enzyme (e.g. Nt.BspQ1) and polymerase extended by nick translation using fluorochrome tagged nucleotides (e.g. Alexa Fluor 647-dUTP). Upon DNA presentation and staining with YOYO-1, DNA molecules are imaged in green (direct excitation of YOYO-1 with laser illumination to detect DNA) and red (FRET with YOYO-1 as donor and Alexa Fluor 647 as acceptor; Alexa Fluor 647 emission; detect nick sites on DNA molecules) channels using appropriate filters. DNA presentation is accomplished in two different ways. First, we use conventional Optical Mapping DNA presentation scheme, where nanocoded DNA is elongated and immobilized on charged glass surfaces *via* flow in microchannels and imaged. Alternatively, we use DNA nanoconfinement in nanoslits to achieve

highly consistent DNA elongation (Kim et al. 2011; Kounovsky-Shafer et al. 2013). During nanoconfinement, DNA molecules can be consistently stretched out very close to their contour lengths, which allows closely spaced genomic markers to be resolved using fluorescence microscopy and image analysis. Additionally, single molecules of DNA are presented in single nanoslits, which eliminate the possibility of experimental artifacts and other chimeras resulting from overlapping DNA molecules. For the analysis of imaging data, superimposed images from green and red channels are processed and analyzed with a newly developed machine vision software to generate Nmap datasets for iterative assembly (INCA; developed Dr. Prabu Ravindran; unpublished).

Nanocoding has resulted from significant advances in physical markup of genomic DNA molecules and subsequent DNA presentation and image processing schemes. These advances have enabled a new high-throughput and low-error system, which promises to generate a better understanding of genome structure in normal and cancer genomes.

1.4. Genomic aberrations in multiple myeloma

Multiple Myeloma is a B-cell malignancy characterized by slow proliferation of terminally differentiated, antibody producing plasma cells within the bone marrow (Palumbo & Anderson 2011). It is generally preceded by a benign premalignant disorder, monoclonal gammopathy of undetermined significance (MGUS) and in some patients, followed by a more aggressive form of tumor, which proliferates in extramedullary sites and is termed plasma cell leukemia (PCL). It is a relatively uncommon cancer, with a 0.7% (1 in 143) lifetime risk in the United States. The American Cancer Society estimates 24,050 new cases and 11,090 deaths

related to multiple myeloma in the United States for the year 2014 (http://www.cancer.org/cancer/multiplemyeloma/).

Like other cancers, multiple myeloma has genomic underpinnings and is characterized by many genomic aberrations, ranging from point mutations to aneuploidy. With technological improvements that have transformed our ability to study human genomes over time, our understanding of these aberrations in multiple myeloma has also improved significantly.

1.4.1. Multiple myeloma genomics using conventional approaches

Traditionally, cytogenetic banding analysis was used to study multiple myeloma patient samples and cell lines. Using such methods, numerical and structural aberrations were observed in approximately 30 - 50% patient samples where metaphase chromosomal spreads could be achieved (Sawyer et al. 1995; Calasanz et al. 1997). These aberrations included trisomies of many odd-numbered chromosomes (excluding chr13) and translocations, predominantly involving the immunoglobulin heavy chain (IGH) locus on chr14. However, banding analysis underestimated genomic irregularities in multiple myeloma because of three reasons. Firstly, plasma cells have a low mitotic index (Drewinko et al. 1981) and consequently, do not always proliferate to give high quality metaphase cells, which are required for high quality banding analysis. Secondly, in culture, normal cells tend to proliferate preferentially and present normal karyotypes, which are not reflective of the disease. Finally, banding methods have low resolution (~10 Mb; Shaffer & Bejjani 2004) and hence, submicroscopic aberrations including interstitial deletions, duplications and other rearrangements are not captured in banding analysis.

As discussed earlier, fluorescence in situ hybridization (FISH) (Gray & Pinkel 1992) uses interphase chromosomes and hence, it can be used to overcome some of the shortcomings of banding analysis. When FISH was used for multiple myeloma, >90% of all patient samples presented numerical and structural aberrations, highlighting the presence of genomic aberrations in almost all multiple myeloma cases (Liebisch et al. 2003). However, like previously discussed, FISH based methods are limited in their ability to discover genomic variation.

Comparative Genome Hybridization (CGH) (Cigudosa et al. 1998) and later, array CGH (aCGH) and SNP genotyping methods (Walker et al. 2006; Avet-Loiseau et al. 2009; Walker et al. 2010) were used to assay copy number changes across many multiple myeloma patient genomes. These methods, which are based on the differences in fluorescence intensity observed when differentially labeled tumor and control DNA samples are hybridized to a normal DNA sample, improved the resolution of detectable copy number changes to a few kilobases and also allowed assays of loss of heterozygosity. Common copy number aberrations were found at 1p, 6q, 8p, 12p, 13q, 14q, 16q, 17p, 20, and 22 (losses) and at 1q and 6p (gains). Many of these aberrations like 1p loss, 1q gain and 17p loss were found to have adverse effect on overall survival (Walker et al. 2010). These mapping methods, although global in scale and high in resolution, have limitations of their own. They cannot identify balanced events like inversions, translocations and other complex rearrangements and also lack the details of structural changes that accompany copy number changes. Furthermore, they lack the resolution to identify smaller mutational events, particularly single nucleotide variants, indels and structural variants below their resolution, which have since been shown to be important drivers of multiple myeloma pathogenesis (Chapman MA et al. 2011; Lohr et al. 2014).

Collectively, these approaches have revealed a general picture of genomic aberrations and their impact on multiple myeloma pathogenesis. The landscape of such aberrations in multiple myeloma lies between hematologic neoplasms, which harbor limited changes, and solid tumors, which present with a wide variety of chromosomal and genomic rearrangements (Morgan et al. 2012). We now know that almost all multiple myeloma patient samples contain genomic aberrations, including many recurrent ones. Accordingly, multiple myeloma was classified into two subtypes: hyperdiploid and non-hyperdiploid (Debes-Marun et al. 2003; Fonseca et al. 2004). Hyperdiploid multiple myeloma genome contains 48 to 74 chromosomes, accounts for approximately 50% of all cases, is characterized by trisomies of many odd-numbered chromosomes (3, 5, 7, 9, 11, 15, 19 and 21; excluding chr13), and low prevalence of translocations at the immunoglobulin heavy chain (IGH) locus on chr14 and deletions on chr13. On the other hand, the non-hyperdiploid multiple myeloma genome contains less than 48 or greater than 74 chromosomes, is characterized by translocations involving the immunoglobulin heavy chain locus on chr14 and is often associated with hemizygous loss of all or part of chr13 (Fonseca et al. 2003). These aberrations are not exclusive to the defined subtypes because overlaps are commonly found between the two. These subtypes have an impact on clinical outcomes: hyperdiploidy is associated with a more favorable outcome when compared to non-hyperdiploid cases, which are generally associated with an adverse outcome. Some aberrations like deletions of 13q are associated with a significantly shorter overall survival (Schmidt-Wolf et al. 2006). The underlying genetic aberrations have also been shown to impact drug response. For instance, induction therapy or long term treatment with bortezomib has been shown to overcome the poor prognosis associated with t(4;14) (San Miguel et al. 2008; Pineda-Roman et

al. 2008). Many of these aberrations cannot explain malignant transformation because they are also observed in similar proportions in pre-malignant MGUS patients (Fonseca et al. 2002). Others like MYC abnormalities, RAS and TP53 mutations distinguish multiple myeloma from MGUS (Liu et al. 1996; Bezieau et al. 2001).

1.4.2. Multiple myeloma genomics using DNA sequencing

With the technological advances made in next generation sequencing platforms (Metzker 2010), a deeper appreciation of genome variation in multiple myeloma has emerged over the last 5 years. Most importantly, sequencing approaches have enabled querying the multiple myeloma genome at single nucleotide resolution, and consequently, revealed the patterns of single nucleotide variation in multiple myeloma (Chapman MA et al. 2011; Lohr et al. 2014). Recent studies have demonstrated widespread inter-tumor (Chapman MA et al. 2011) and intra-tumor (Lohr et al. 2014; Bolli et al. 2014) genetic heterogeneity. They have also uncovered the patterns of clonal evolution (Keats et al. 2012; Egan et al. 2012; Bolli et al. 2014) including the presence of clonal tides (Egan et al. 2012) and alternating clonal dominance (Keats et al. 2012) in multiple myeloma.

Chapman and colleagues, in their pivotal study describing sequencing of 38 multiple myeloma patient samples, discovered widespread inter-tumor heterogeneity at the level of single nucleotide variation (Chapman MA et al. 2011). They identified that apart from commonly mutated NRAS, KRAS and TP53, genes involved in protein translation, histone methylation and blood coagulation were significantly mutated in multiple myeloma. They found, on average, 21 chromosomal rearrangements disrupting protein coding regions in each patient sample. Although individual mutations varied, the authors suggested that common

downstream pathways are affected as a result of these seemingly disparate mutations. They also found potential clinically actionable mutations in BRAF in a small subset of patients. A larger follow-up study (Lohr et al. 2014) involving 203 patient samples identified 11 significantly mutated genes in multiple myeloma. A majority of patient samples were shown to have one or more subclones, indicating intra-tumor genomic heterogeneity. However, only 82% of the patients harbored recurrent single nucleotide mutations, suggesting that other genetic, epigenetic or transcriptomic factors might contribute to multiple myeloma pathogenesis in the remaining patients.

Since 2012, a number of studies have described intra-clonal architectural heterogeneity at different stages of multiple myeloma progression (Egan et al. 2012; Keats et al. 2012; Bolli et al. 2014; Melchor et al. 2014). Longitudinal analysis of a multiple myeloma genome from diagnosis to leukemic transformation in a t(4;14) patient revealed genomic evolution and clonal tides based on single nucleotide variants (SNVs) detected at different stages of disease progression (Egan et al. 2012). Of the 27 SNVs that the authors found to be located in genes known to harbor mutations in MMRC cohort (Chapman MA et al. 2011) and/or COSMIC database, 10 were shared among all the samples while 5 were seen only in the last stage PCL sample and possibly contributed to leukemic transformation. Another similar longitudinal study analyzed a multiple myeloma patient with 1q amplification at disease presentation and relapse, and found an increase in mutational complexity with disease progression (Weston-Bell et al. 2013). The authors identified 81 non-synonymous variants, with 33 shared at presentation and relapse and another 48 novel to just the relapse sample. Serial analysis of copy number changes in 28 multiple myeloma patients using genome CGH arrays revealed heterogeneity in patterns

of genome evolution (Keats et al. 2012). The authors identified three phenotypic patterns of genome evolution: no change, linear evolution and heterogeneous clonal mixtures with shifting predominant clones. These findings are also clinically important; the authors suggested that poor prognosis of cytogenetically high-risk multiple myeloma might be related to increased clonal heterogeneity and genomic instability. These findings have recently been corroborated by single cell analysis based on single nucleotide variation, which identified linear, branching and parallel modes of clonal evolution in multiple myeloma (Melchor et al. 2014). Connecting copy number changes and single nucleotide variation, a recent study suggested that aneuploidy and point mutations act in concert over time and across subclones (Bolli et al. 2014). Previous reports have associated point mutations in DIS3, one of the significantly mutated genes from MMRF cohort, to be associated with chr13 deletion (Walker et al. 2012).

1.5. Motivation behind our work and summary of thesis

Putting it all together, it is now well established that the multiple myeloma genome mutates in profound ways, which contributes to pathogenesis and tumor evolution, and affects clinical outcomes including treatment response, chemotherapeutic resistance and overall survival. While consensus has started to emerge about point mutations and large-scale copy number changes, which lie on the ends of the size spectrum of mutational changes, we still don't understand much about structural variation in multiple myeloma. Also, no single gene mutation or combination of mutations has emerged as universal to multiple myeloma genomes at presentation, highlighting that aberrant molecular pathways leading to tumorigenesis are varied and a comprehensive understanding of the patterns of this variation might lead to improved clinical outcomes following genome guided personalized therapy. Although a few

studies (Chapman MA et al. 2011; Egan et al. 2012; Lohr et al. 2014) mention the identification of structural variants in the analyzed multiple myeloma genomes, a rigorous approach to determination and analysis of structural variation is still missing. A better understanding of structural variation in conjunction with single nucleotide and copy number variation might improve our understanding of the underlying molecular processes in multiple myeloma.

Additionally, except for a few recent studies (Egan et al. 2012; Bolli et al. 2014), most of the research towards understanding of multiple myeloma genomes has focused on studying genomic changes at a single timepoint (presentation or after treatment etc.). By identifying and studying genomic changes associated with clinical transformation from a drug sensitive to a refractory stage, we can generate functional hypothesis for molecular pathways that might contribute to drug resistance. These could then be studied for their functional contribution to drug resistance and utilized to define further treatments.

To address these issues, we designed a study to perform serial analysis of tumor samples from two time-points during disease progression in a multiple myeloma patient: a drug sensitive state (MM-S) and a drug refractory state (MM-R), with the objective of comprehensively identifying somatic variants at these time-points by comparing them to a normal sample obtained from the same patient (Normal). Our lab, the Laboratory for Molecular and Computation Genomics (LMCG), has developed Optical Mapping for physical mapping of whole genomes. As discussed earlier, Optical Mapping addresses many of the limitations with existing genomic analysis systems. In chapter 2, I will discuss the application of Optical Mapping towards characterizing structural and copy number variation, and delineating genomic structure and rearrangements in these multiple myeloma tumor samples. In chapter 3, I will

discuss the application of paired-end DNA sequencing to identify somatic variation in these samples. By integrating complementary findings from Optical Mapping and DNA sequencing, we have comprehensively identified genomic variation, ranging from single nucleotide variants to chromosomal gains and losses in these samples, which I will present in chapter 3. In chapter 4, I will conclude this thesis with a brief discussion and directions for future.

1.6. Bibliography

1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–73.

Abecasis, G.R. et al., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65.

Abel, H.J. & Duncavage, E.J., 2013. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer genetics*, 206(12), pp.432–40.

Abyzov, A. et al., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6), pp.974–84.

Albertson, D.G., 2003. Profiling breast cancer by array CGH. *Breast cancer research and treatment*, 78(3), pp.289–98.

Alexandrov, L.B. et al., 2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415–21.

Alkan, C., Coe, B.P. & Eichler, E.E., 2011. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5), pp.363–76.

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

Anantharaman, T., Mishra, B. & Schwartz, D., 1999. Genomics via optical mapping. III: Contiging genomic DNA. *International Conference on Intelligent Systems for Molecular Biology*, pp.18–27.

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

Antonacci, F. et al., 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature genetics*, 42(9), pp.745–50.

Armbrust, E.V. et al., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, 306(5693), pp.79–86.

Avery, O.T., Macleod, C.M. & McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolation from pneumococcus Type III. *The Journal of experimental medicine*, 79(2), pp.137–58.

Avet-Loiseau, H. et al., 2009. Prognostic significance of copy-number alterations in multiple myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(27), pp.4585–90.

Bauman, J.G. et al., 1980. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. *Experimental cell research*, 128(2), pp.485–90.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–9.

Benvenuti, S. et al., 2007. Oncogenic activation of the RAS/RAF signaling pathway impairs the response of metastatic colorectal cancers to anti-epidermal growth factor receptor antibody therapies. *Cancer research*, 67(6), pp.2643–8.

Bezieau, S. et al., 2001. High incidence of N and K-Ras activating mutations in multiple myeloma and primary plasma cell leukemia at diagnosis. *Human mutation*, 18(3), pp.212–24.

Biankin, A. V et al., 2012. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, 491(7424), pp.399–405.

Bignell, G.R. et al., 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome research*, 14(2), pp.287–95.

Bolli, N. et al., 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5, p.2997.

Boveri, T., 2008. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of cell science*, 121 Suppl , pp.1–84.

Cahill, D.P. et al., 2007. Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(7), pp.2038–45.

Cai, W. et al., 1998. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 95(7), pp.3390–5.

Cai, W. et al., 1995. 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), pp.5164–8.

Calasanz, J. et al., 1997. Cytogenetic Analysis of 280 Patients With Multiple Myeloma and Related Disorders : Primary Breakpoints. , 93(June 1996), pp.84–93.

Campbell, P.J. et al., 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6), pp.722–9.

Campbell, P.J. et al., 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319), pp.1109–13.

Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), pp.1061–8.

Cancer Genome Atlas Research Network, 2012a. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417), pp.519–25.

Cancer Genome Atlas Research Network, 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456), pp.43–9.

Cancer Genome Atlas Research Network, 2012b. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–7.

Cancer Genome Atlas Research Network, 2014. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492), pp.315–22.

Cancer Genome Atlas Research Network, 2012c. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70.

Cancer Genome Atlas Research Network, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), pp.609–15.

Caspersson, T. et al., 1968. Chemical differentiation along metaphase chromosomes. *Experimental cell research*, 49(1), pp.219–22.

Caspersson, T., Zech, L. & Johansson, C., 1970. Differential binding of alkylating fluorochromes in human chromosomes. *Experimental cell research*, 60(3), pp.315–9.

Chapman, M.A. et al., 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), pp.467–72.

Chapman, P.B. et al., 2011. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England journal of medicine*, 364(26), pp.2507–16.

Chen, K. et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9), pp.677–81.

Chen, S. et al., 2012. Genome sequence of the model medicinal mushroom Ganoderma lucidum. *Nature communications*, 3, p.913.

Cigudosa, J.C. et al., 1998. Characterization of nonrandom chromosomal gains and losses in multiple myeloma by comparative genomic hybridization. *Blood*, 91(8), pp.3007–10.

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

Conrad, D.F. et al., 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nature genetics*, 38(1), pp.75–81.

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

Debes-Marun, C.S. et al., 2003. Chromosome abnormalities clustering and its implications for pathogenesis and prognosis in myeloma. *Leukemia*, 17(2), pp.427–36.

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

Ding, L. et al., 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382), pp.506–10.

Ding, L. et al., 2010. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291), pp.999–1005.

Ding, L. et al., 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216), pp.1069–75.

Drewinko, B. et al., 1981. The growth fraction of human myeloma cells. *Blood*, 57(2), pp.333–8.

Druker, B.J., 2004. Imatinib as a paradigm of targeted therapies. *Advances in cancer research*, 91, pp.1–30.

Druker, B.J., 2008. Translation of the Philadelphia chromosome into therapy for CML. *Blood*, 112(13), pp.4808–17.

Egan, J.B. et al., 2012. Whole genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution and clonal tides. *Blood*. 120(5), pp.1060-6.

Esteva, F.J. et al., 2010. Molecular predictors of response to trastuzumab and lapatinib in breast cancer. *Nature reviews. Clinical oncology*, 7(2), pp.98–107.

Fonseca, R. et al., 2004. Genetics and cytogenetics of multiple myeloma: a workshop report. *Cancer research*, 64(4), pp.1546–58.

Fonseca, R. et al., 2002. Genomic abnormalities in monoclonal gammopathy of undetermined significance. *Blood*, 100(4), pp.1417–24.

Fonseca, R. et al., 2003. The recurrent IgH translocations are highly associated with nonhyperdiploid variant multiple myeloma. *Blood*, 102(7), pp.2562–7.

Francke, U., 1976. Retinoblastoma and chromosome 13. *Cytogenetics and cell genetics*, 16(1-5), pp.131–4.

Gall, J.G. & Pardue, M.L., 1969. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, 63(2), pp.378–83.

Girirajan, S. et al., 2011. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS genetics*, 7(11), p.e1002334.

Govindan, R. et al., 2012. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6), pp.1121–34.

Gray, J.W. & Pinkel, D., 1992. Molecular cytogenetics in human cancer diagnosis. *Cancer*, 69(6 Suppl), pp.1536–42.

Greenman, C. et al., 2007. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132), pp.153–8.

Groffen, J. et al., 1984. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell*, 36(1), pp.93–9.

Harbour, J.W. et al., 2010. Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science (New York, N.Y.)*, 330(6009), pp.1410–3.

Hartwell, L.H. & Lander, E.S., 2005. Report to National Cancer Advisory Board: NCAB Working Group on Biomedical Technology. Available at: http://cancergenome.nih.gov/PublishedContent/Files/pdfs/6.5.1.4_NCABReport_Feb05.pdf [Accessed December 29, 2014b].

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

Heisterkamp, N. et al., Structural organization of the bcr gene and its role in the Ph' translocation. *Nature*, 315(6022), pp.758–61.

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

Hopman, A.H. et al., 1986. Bi-color detection of two target DNAs by non-radioactive in situ hybridization. *Histochemistry*, 85(1), pp.1–4.

Hu, X., 1997. Development of optical primer extension (OPE), and, Improvement and characterization of the optical mapping system. Ph.D. thesis, New York University.

Hunter, C. et al., 2006. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer research*, 66(8), pp.3987–91.

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

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–45.

Jackson, D.A., Symons, R.H. & Berg, P., 1972. 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), pp.2904–9.

Jing, J. et al., 1998. 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), pp.8046–51.

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

Jo, K. et al., 2007. 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), pp.2673–8.

Jo, K., Schramm, T.M. & Schwartz, D.C., 2009. A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding. *Methods in molecular biology (Clifton, N.J.)*, 544, pp.29–42.

Jones, P.A. & Baylin, S.B., 2007. The epigenomics of cancer. *Cell*, 128(4), pp.683–92.

Jones, S. et al., 2008. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11), pp.4283–8.

Kallioniemi, A. et al., 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, 258(5083), pp.818–21.

Kandoth, C., Schultz, N., et al., 2013. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447), pp.67–73.

Kandoth, C., McLellan, M.D., et al., 2013. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471), pp.333–9.

Keats, J.J. et al., 2012. Clonal competition with alternating dominance in multiple myeloma. *Blood*. 120(5), pp.1067-76.

Kelly, T.J. & Smith, H.O., 1970. A restriction enzyme from Hemophilus influenzae. II. *Journal of molecular biology*, 51(2), pp.393–409.

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

Kim, Y. et al., 2011. Nanochannel confinement: DNA stretch approaching full contour length. *Lab on a chip*, 11(10), pp.1721–9.

De Klein, A. et al., 1982. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, 300(5894), pp.765–7.

Kobayashi, S. et al., 2005. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *The New England journal of medicine*, 352(8), pp.786–92.

Korbel, J.O. et al., 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849), pp.420–6.

Kounovsky-Shafer, K.L. et al., 2013. Presentation of large DNA molecules for analysis as nanoconfined dumbbells. *Macromolecules*, 46(20), pp.8356–8368.

Kovacs, G., 1978. Abnormalities of chromosome No. 1 in human solid malignant tumours. *International journal of cancer. Journal international du cancer*, 21(6), pp.688–94.

Kwak, E.L. et al., 2010. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *The New England journal of medicine*, 363(18), pp.1693–703.

De la Chapelle, A., Schröder, J. & Vuopio, P., 1972. 8-Trisomy in the bone marrow. Report of two cases. *Clinical genetics*, 3(6), pp.470–6.

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

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

Lee, W. et al., 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297), pp.473–7.

Ley, T.J. et al., 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218), pp.66–72.

Lichter, P. et al., 1988. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Human genetics*, 80(3), pp.224–34.

Liebisch, P. et al., 2003. Value of comparative genomic hybridization and fluorescence in situ hybridization for molecular diagnostics in multiple myeloma. *British journal of haematology*, 122(2), pp.193–201.

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

Limon, J., Dal Cin, P. & Sandberg, A.A., 1986. Translocations involving the X chromosome in solid tumors: presentation of two sarcomas with t(X;18)(q13;p11). *Cancer genetics and cytogenetics*, 23(1), pp.87–91.

Lin, J. et al., 1999. Whole-genome shotgun optical mapping of Deinococcus radiodurans. *Science (New York, N.Y.)*, 285(5433), pp.1558–62.

Link, A.J. & Olson, M. V, 1991. Physical map of the Saccharomyces cerevisiae genome at 110-kilobase resolution. *Genetics*, 127(4), pp.681–98.

Liu, B.P. et al., 1996. Activating Mutations of N- and. K-ras in multiple myeloma show different clinical associations: analysis of the Eastern Cooperative Oncology Group Phase III Trial. *Blood*, 88(7), pp.2699–2706.

Lobban, P.E. & Kaiser, A.D., 1973. Enzymatic end-to end joining of DNA molecules. *Journal of molecular biology*, 78(3), pp.453–71.

Lohr, J.G. et al., 2014. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell*, 25(1), pp.91–101.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.

Maher, C.A. et al., 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234), pp.97–101.

Manolov, G. & Manolova, Y., 1972. Marker band in one chromosome 14 from Burkitt lymphomas. *Nature*, 237(5349), pp.33–4.

Mardis, E.R. et al., 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England journal of medicine*, 361(11), pp.1058–66.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80.

McDonald, J.C. et al., 2000. Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis*, 21(1), pp.27–40.

Melchor, L. et al., 2014. Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. *Leukemia*, 28(8), pp.1705–15.

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

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp.31–46.

Mills, R.E. et al., 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, 16(9), pp.1182–90.

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

Mitelman, F., Johansson, B. & Mertens, F., 2004. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature genetics*, 36(4), pp.331–4.

Mitelman, F., Johansson, B. & Mertens, F., 2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer*, 7(4), pp.233–45.

Mitsudomi, T. et al., 2010. Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *The Lancet. Oncology*, 11(2), pp.121–8.

Moorhead, P.S. et al., 1960. Chromosome preparations of leukocytes cultured from human peripheral blood. *Experimental cell research*, 20, pp.613–6.

Morgan, G.J., Walker, B. a & Davies, F.E., 2012. The genetic architecture of multiple myeloma. *Nature reviews. Cancer*, 12(5), pp.335–48.

Muleris, M. et al., 1985. Consistent deficiencies of chromosome 18 and of the short arm of chromosome 17 in eleven cases of human large bowel cancer: a possible recessive determinism. *Annales de génétique*, 28(4), pp.206–13.

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

Ng, J.M.K. et al., 2002. Components for integrated poly(dimethylsiloxane) microfluidic systems. *Electrophoresis*, 23(20), pp.3461–73.

Nik-Zainal, S. et al., 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5), pp.979–93.

Nowell, P.C. & Hungerford, D.A., 1960. Chromosome studies on normal and leukemic human leukocytes. *Journal of the National Cancer Institute*, 25, pp.85–109.

Palumbo, A. & Anderson, K., 2011. Multiple myeloma. *The New England journal of medicine*, 364(11), pp.1046–60.

Pao, W. et al., 2005. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS medicine*, 2(3), p.e73.

Pardue, M.L. & Gall, J.G., 1969. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, 64(2), pp.600–4.

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

Pineda-Roman, M. et al., 2008. Sustained complete remissions in multiple myeloma linked to bortezomib in total therapy 3: comparison with total therapy 2. *British journal of haematology*, 140(6), pp.625–34.

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

Pinkel, D. & Albertson, D.G., 2005. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37 Suppl, pp.S11–7.

Pleasance, E.D., Cheetham, R.K., et al., 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278), pp.191–6.

Pleasance, E.D., Stephens, P.J., et al., 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278), pp.184–90.

Ray, M. et al., 2013. Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis. *BMC genomics*, 14, p.505.

Reddy, E.P. et al., 1982. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888), pp.149–52.

Reslewic, S. et al., 2005. Whole-genome shotgun optical mapping of Rhodospirillum rubrum. *Applied and environmental microbiology*, 71(9), pp.5511–22.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–52.

Rowley, J.D., 1973a. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Annales de génétique*, 16(2), pp.109–12.

Rowley, J.D., 1973b. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243(5405), pp.290–3.

San Miguel, J.F. et al., 2008. Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma. *The New England journal of medicine*, 359(9), pp.906–17.

Sawyer, J.R. et al., 1995. Cytogenetic findings in 200 patients with multiple myeloma. *Cancer genetics and cytogenetics*, 82(1), pp.41–9.

Schmidt-Wolf, I.G.H. et al., 2006. Chromosomal aberrations in 130 patients with multiple myeloma studied by interphase FISH: diagnostic and prognostic relevance. *Cancer genetics and cytogenetics*, 167(1), pp.20–5.

Schröck, E. et al., 1996. Multicolor spectral karyotyping of human chromosomes. *Science (New York, N.Y.)*, 273(5274), pp.494–7.

Schwaenen, C. et al., 2004. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4), pp.1039–44.

Schwartz, D.C. et al., 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (New York, N.Y.)*, 262(5130), pp.110–4.

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

Sebat, J. et al., 2004. Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305(5683), pp.525–8.

Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823), pp.445–9.

Shaffer, L.G. & Bejjani, B.A., A cytogeneticist's perspective on genomic microarrays. *Human reproduction update*, 10(3), pp.221–6.

Shah, S.P. et al., 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265), pp.809–13.

Simpson, J.T. et al., 2009. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6), pp.1117–23.

Skiadas, J. et al., 1999. Optical PCR: genomic analysis by long-range PCR and optical mapping. *Mammalian genome : official journal of the International Mammalian Genome Society*, 10(10), pp.1005–9.

Smith, H.O. & Wilcox, K.W., 1970. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *Journal of molecular biology*, 51(2), pp.379–91.

Soda, M. et al., 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153), pp.561–6.

Solinas-Toldo, S. et al., 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, chromosomes & cancer*, 20(4), pp.399–407.

Stehelin, D. et al., 1976. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547), pp.170–3.

Stephens, P.J. et al., 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276), pp.1005–10.

Stephens, P.J. et al., 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1), pp.27–40.

Stratton, M.R., 2011. Exploring the genomes of cancer cells: progress and promise. *Science (New York, N.Y.)*, 331(6024), pp.1553–8.

Stratton, M.R., Campbell, P.J. & Futreal, P.A., 2009. The cancer genome. *Nature*, 458(7239), pp.719–24.

Sudmant, P.H. et al., 2010. Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, 330(6004), pp.641–6.

Teague, B. et al., 2010. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24), pp.10848–53.

Tjio, J.H. & Levan, A., 2010. The chromosome number of man. *Hereditas*, 42(1-2), pp.1–6.

Trask, B.J., 2002. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature reviews. Genetics*, 3(10), pp.769–78.

Turcatti, G. et al., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4), p.e25.

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

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

Valouev, A., Schwartz, D.C., et al., 2006. 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), pp.15770–5.

Valouev, A., Zhang, Y., et al., 2006. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, 22(10), pp.1217–24.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.

Vogelstein, B. et al., 2013. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127), pp.1546–58.

Vorsanova, S.G., Yurov, Y.B. & Iourov, I.Y., 2010. Human interphase chromosomes: a review of available molecular cytogenetic technologies. *Molecular cytogenetics*, 3, p.1.

Walker, B. a et al., 2010. A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, 116(15), pp.e56–65.

Walker, B. a et al., 2006. Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. *Blood*, 108(5), pp.1733–43.

Walker, B.A. et al., 2012. Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma. *Blood*, 120(5), pp.1077–86.

Watson, J.D. & Crick, F.H.C., 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356), pp.737–738.

Wei, F. et al., 2009. The physical and genetic framework of the maize B73 genome. *PLoS genetics*, 5(11), p.e1000715.

Weinstein, J.N. et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–20.

Weiss, B. & Richardson, C.C., 1967. Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proceedings of the National Academy of Sciences of the United States of America*, 57(4), pp.1021–8.

Weston-Bell, N. et al., 2013. Exome sequencing in tracking clonal evolution in multiple myeloma following therapy. *Leukemia*, 27(5), pp.1188–91.

Whitesides, G.M. et al., 2001. Soft lithography in biology and biochemistry. *Annual review of biomedical engineering*, 3, pp.335–73.

Wong, K.M., Hudson, T.J. & McPherson, J.D., 2011. Unraveling the genetics of cancer: genome sequencing and beyond. *Annual review of genomics and human genetics*, 12, pp.407–30.

Wood, L.D. et al., 2007. The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, 318(5853), pp.1108–13.

Wu, M. et al., 2008. Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. *Carcinogenesis*, 29(9), pp.1710–6.

Wu, T. & Schwartz, D.C., 2007. Transchip: single-molecule detection of transcriptional elongation complexes. *Analytical biochemistry*, 361(1), pp.31–46.

Ye, K. et al., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21), pp.2865–71.

Zhang, F. et al., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10, pp.451–81.

Zhou, S. et al., 2009. A single molecule scaffold for the maize genome. *PLoS genetics*, 5(11), p.e1000711.

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

Zhou, S. et al., 2004. Single-molecule approach to bacterial genomic comparisons via optical mapping. *Journal of bacteriology*, 186(22), pp.7773–82.

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

Zhou, S. et al., 2003. 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), pp.2142–51.

1.7. Figures

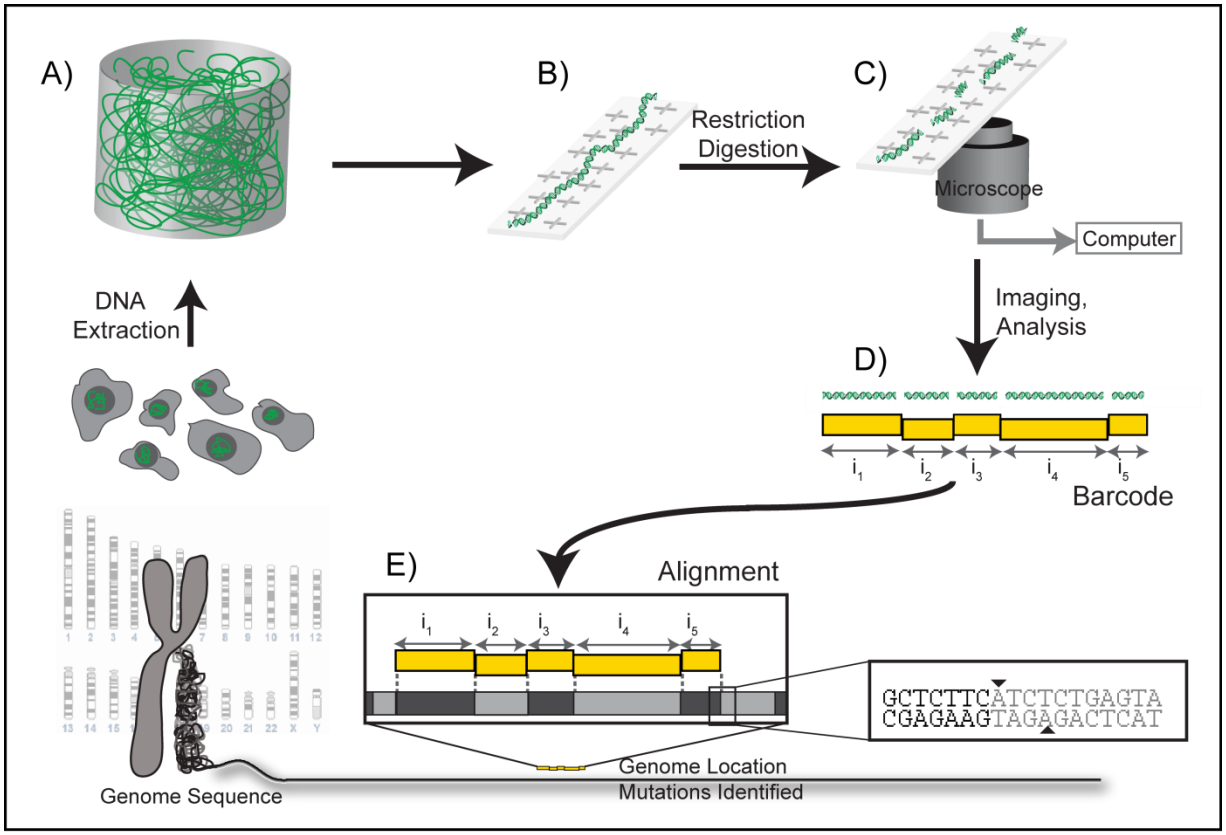

Fig 1-1: An overview of Optical Mapping. High molecular weight genomic DNA is extracted (A) and immobilized on positively charged surfaces in elongated conformation using microchannels (B). It is restriction digested, stained with fluorescent dye YOYO-1 and imaged using automated epifluorescence microscopy workstations (C). The images are processed using custom image processing software to generate single molecule ordered restriction maps (D). Millions of single molecule maps are collected, and then assembled using an iterative assembly algorithm to identify structural variation across the genome (E) (Figure credits: Kristy Kounovsky-Shafer).

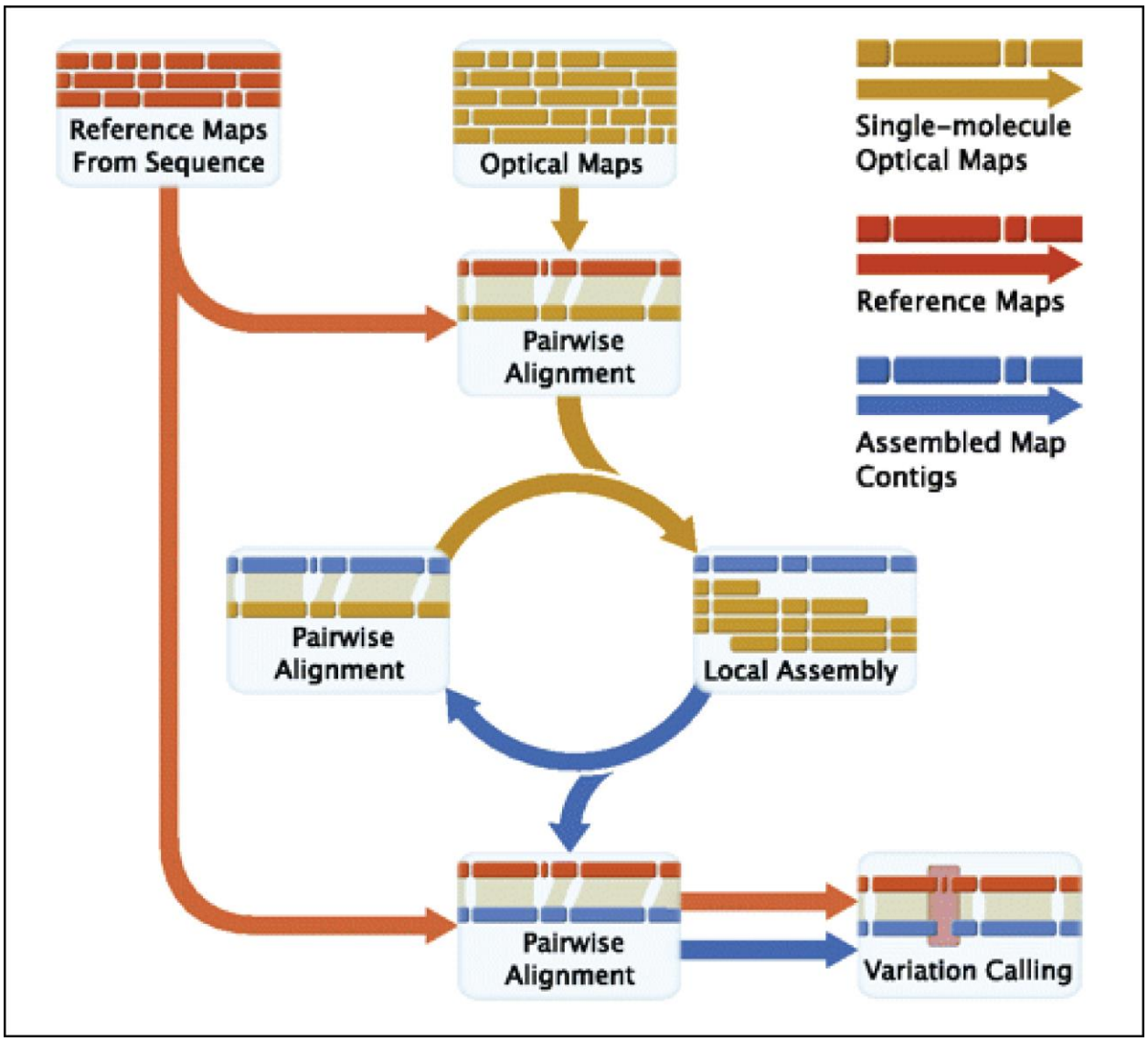

Fig. 1-2: An overview of computational pipeline used for iterative assembly and structural variation calling. Reference maps, generated from *in silico* digestion of human genome reference sequence, are used to seed an iterative assembly process that utilizes pairwise alignment (to cluster together similar single molecule Rmaps) and local assembly (to generate consensus maps from clusters of Rmaps). After several iterations of alignment and local assembly, the final consensus maps are aligned back to the reference maps to identify locations where these maps differ in structure (Figure reproduced from Teague et al. 2010).

Chapter 2: Optical Mapping identifies widespread structural and copy number variation in multiple myeloma

In this chapter, I will describe our use of Optical Mapping for the analysis of structural and copy number variation in a multiple myeloma patient genome at two time-points: a drug sensitive stage (MM-S) and a drug refractory stage (MM-R). Finally, I will present and discuss our findings from this analysis.

2.1. Introduction

It is now well accepted that cancer genomes harbor variation that lies along the entire length spectrum (Mitelman et al. 2007; Ley et al. 2008; Stephens et al. 2009; Weinstein et al. 2013; Kandoth et al. 2013). These variants include single nucleotide changes, indels, structural variants, copy number changes, genomic rearrangements and aneuploidy. Although recent progress in DNA sequencing methods (Margulies et al. 2005; Bentley et al. 2008; Rothberg et al. 2011) has potentiated quick analysis of variation in many cancer genomes, such analysis is limited in scope because of three reasons. First, current sequencing approaches generate short sequencing reads, which cannot be used to unambiguously study low-complexity, repeat-rich regions of the genome known to preferentially harbor structural variants and genomic rearrangements (Korbel et al. 2007; Sudmant et al. 2010). Second, most of the current cancer genome analysis is focused on single nucleotide changes, small indels and copy number changes, which are easily assayed, analyzed and integrated in population studies. Consequently, comprehensive knowledge of cancer genome variation still eludes us because of the technological challenges that complicate structural variation analysis and because of the overwhelming focus on analyzing simple smaller variants. Third and last, most of the current

analysis is based on comparison to the human reference sequence. Although useful, this limits variation calling to our knowledge about the reference genome, which is still incomplete.

Our lab developed Optical Mapping, a high-throughput, single-molecule system, which is used to address most of these limitations, thereby potentiating the analysis of almost comprehensive genome structure, structural variation and rearrangements in complex cancer genomes (Ray et al. 2013). Optical Mapping generates ordered restriction maps (Rmaps) from high molecular weight genomic DNA molecules, ranging in size from 300 kilobases to megabases (Schwartz et al. 1993; Meng et al. 1995; Cai et al. 1995; Cai et al. 1998; Lai et al. 1999; Anantharaman et al. 1999; Dimalanta et al. 2004; Valouev, Schwartz, et al. 2006; Zhou et al. 2007; Teague et al. 2010; Ray et al. 2013). The Rmaps are then assembled using an iterative assembler that generates genome-wide physical maps (Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Teague et al. 2010), which provide insights into genome structure. In chapter 1, I provided a detailed introduction to the Optical Mapping system. Because of its ability to analyze long DNA analytes without the need for amplification and cloning, Optical Mapping allows us to study long-range genome structure and thus, unambiguously identifies structural variants and genomic rearrangements even in low-complexity, repeat-rich regions. Also, physical genome assemblies traverse almost the entire human reference sequence and even many sequence gaps (Teague et al. 2010), thereby generating an almost comprehensive view of genome variation. Additionally, the iterative nature of computational assembly pipeline enables the assembly of highly complex genomic rearrangements that deviate substantially from the reference. In this chapter and the next one, I will leverage all of these unique advantages of Optical Mapping to generate a detailed picture

of structural variation and genomic rearrangements in a multiple myeloma patient genome. This analysis identifies widespread structural variation, which would have been otherwise invisible to sequencing based genome analysis methods.

2.2. Study Design

Using Optical Mapping, we have analyzed two multiple myeloma tumor samples and a matched normal sample. The tumor samples were collected from a multiple myeloma patient at two time-points: when the patient was sensitive to chemotherapeutic treatments (MM-S) and later when the patient became refractory to all treatments (MM-R). The objective of our work was to reconstruct genome structure and identify structural variation, copy number variation and genomic rearrangements in these tumor samples using Optical Mapping (Fig. 2-1).

2.3. Materials and Methods

2. 3.1. Ethics Statement

This study was approved by the Institutional Review Board at the University of Wisconsin-Madison in accordance with the Declaration of Helsinki.

2. 3.2. Patient History and Sample Selection

A 58 year old male patient presented in the clinic with renal failure and hypercalcemia in April 2006 and was diagnosed with ISS Stage IIIb multiple myeloma, with multiple lytic bone lesions and 63% plasma cells in bone marrow biopsy. The patient was initially treated with bortezomib/dexamethasone followed by tandem autologous stem cell transplants in October 2006 and January 2007. However, the patient relapsed in January 2008 with karyotypic abnormalities including hypoploidy and del(17p), and received salvage therapy with lenalidomide/dexamethasone followed by bortezomib/dexamethasone. Bone marrow sample

from August 2008 (MM-S sample) showed effacement of marrow cavity with plasma cells and multiple new karyotypic abnormalities including structural losses of 1, 2, 7, 13, 14 and 17. At this time, the patient achieved very good partial response (VGPR) to cyclophosphamide, but relapsed again in October 2009, with 90% plasma cells in bone marrow sample (MM-R sample). The patient died of progressive disease in February 2010, approximately 4 years after first diagnosis.

From bone marrow aspirates collected from the patient in August 2008 (MM-S; patient responded to Subsequent treatments) and October 2009 (MM-R; patient was Refractory to all subsequent treatments), CD138+ positive cells were isolated using AutoMACS Pro Miltenyi Separation System as described before (Kim et al. 2012). Mesenchymal stromal cells, cultured from MM-R sample (Passage 3), were used as the matched Normal sample (MM-Normal). These samples were collected by Dr. Peiman Hematti and provided to us by Dr. Fotis Asimakopoulos at Wisconsin Institute of Medical Research (WIMR).

2.3.3. Optical Mapping

2.3.3.1. DNA preparation: High molecular weight genomic DNA was extracted using a gentle liquid lysis protocol. Briefly, cell samples were thawed on ice, washed with 1 ml ice-cold DPBS (Gibco) twice and resuspended in 1X TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) with 34 pg/µL Lambda DashII DNA (Stratagene) and 0.5 mg/mL Proteinase K (Bioline USA), at a concentration of 20 to 200 cells/µL. The cell suspension was distributed into 200 µL aliquots and incubated at 72°C for 10 minutes, then left overnight at 25°C and finally stored at 4°C for further use.

2.3.3.2. Cleaning Optical Mapping surfaces: 22 X 22 mm cover glass slips (Fisherfinest, Fisher Scientific) were mounted in a custom polytetrafluoroethylene (PTFE) rack in a sealed reaction

vessel. They were cleaned by heating at 70 °C with Nano-strip (Cyantek) for 1 h followed by multiple rinses with high purity water, then heating at 104 °C with 12.1 N hydrochloric acid (Fisher Scientific) for 6 h, and a final set of rinses with high purity water. The cleaned surfaces were washed and then stored in absolute ethanol (Pharmco-AAPER) until use.

2.3.3.3. Derivatizing Optical Mapping surfaces: The cleaned glass cover slips were loaded on a custom PTFE rack and incubated in 250 mL ddH_2O with 90 µL of N-trimethoxylsilylpropyl-N,N,N-trimethylammonium chloride (50% solution in methanol) and 10 µL of vinyltrimethoxysilane (Gelest) with light agitation at 65 °C for 17.5 h. They were rinsed twice with distilled water, then twice with absolute ethanol and finally stored for use in absolute ethanol. This process imparts positive charge to the cleaned surfaces, which helps in immobilizing DNA in later steps (Cai et al. 1995).

2.3.3.4. DNA deposition and digestion: The Optical Mapping surfaces were taken out of ethanol incubation and left to dry under ambient conditions. Once dried, high molecular weight genomic DNA was elongated on Optical Mapping surfaces *via* capillary flow in a microfluidic device made of poly-dimethylsiloxane with microchannels that were 10 mm long, 100 µm wide and 3 µm high (Dimalanta et al. 2004). DNA is immobilized on positively charged glass surfaces *via* electrostatic interactions and elongated *via* flow-based forces in microchannels (Dimalanta et al. 2004). A 3.3 % polyacrylamide overlay (111 µL 29:1 acrylamide: bisacrylamide solution (Amresco), 889 µL ddH_2O, 7.5 µL 10% ammonium persulfate (APS), 0.8 µL TEMED) was then deposited on the arrays of elongated DNA molecules. The polyacrylamide overlay prevents desorption of small fragments of DNA, which are formed after restriction digestion (Zhou, Deng, et al. 2002). The surfaces were then rinsed twice with 1 mL TE, equilibrated with 200 µL

restriction digestion buffer (1X NEB buffer 3) and finally incubated in a humidified chamber with 20 U of BamHI (New England Biolabs) in 200 µL, 1X NEB buffer 3 solution containing 0.02% Triton-X100 (Roche) at 37 °C. The digestion time varied from 30 minutes to 2 h based on the quality of data obtained in preliminary data collection. After digestion was complete, BamHI solution was washed off with 2 rinses of 1X TE. Restriction fragments from DNA molecules were then stained with DNA intercalating dye YOYO-1 (12 µL 0.5 µM in β-mercaptoethanol/TE; Life Technologies). The restriction enzyme (BamHI) introduces double-stranded DNA cuts at cognate restriction sites on DNA molecules, which leads to small gaps in elongated molecules (~1 µm wide) and coil relaxation at ends of newly formed DNA fragments. The DNA fragments were detected in the next step using epifluorescence microscopy.

2.3.3.5. Data collection via epifluorescence microscopy: The arrays of microchannels were imaged at an Optical Mapping workstation, which comprises a Zeiss 135M inverted microscope (Carl Zeiss, Thornwood, NY) equipped with 63X oil immersion objective and illuminated by 488 nm Argon ion laser (Spectra-Physics). Other components e.g. a high speed CCD camera, shutter, stage and focus controllers are interconnected and controlled to yield an automated image collection and processing system (Dimalanta et al. 2004; Teague et al. 2010). The Optical Mapping surfaces with digested and stained DNA can be imaged automatically using these custom-designed microscopy systems with very little setup time.

2.3.3.6. Automated image analysis to generate ordered restriction maps: As the microchannels were imaged, they were automatically processed with two pieces of software. The first one uses a bright image to flatten the raw micrographs, which eliminates possible errors due to uneven illumination, and then overlaps and merges the series of images collected from each

microchannel into a single composite image, which is then used to generate useful data from DNA molecules scanning multiple frames. The second software, called Pathfinder, identifies single molecules of DNA and then individual restriction fragments. It sizes the fragments based on relative fluorescence intensities, relative to a sizing standard (lambda Dash II), which is mixed with DNA sample and presented on glass surfaces. The restriction pattern and resulting fragment sizes are known for the sizing standard, and this information is used to size the fragments of genomic DNA from the sample being studied. This process generated ordered restriction maps, which were finally assembled to create genome-wide physical restriction maps for the samples.

2.3.4. Iterative optical map assembly

The computational framework for Optical Map assembly has been described in detail previously (Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Teague et al. 2010). Briefly, each stage of this iterative process is similar to sequence assembly and consists of three steps. First, single molecule Rmaps are aligned to an *in silico* restriction map derived from human reference sequence (NCBI Build 37) using an in-house alignment software called Software for Optical Mapping Analysis (SOMA). This step clusters similar maps to the genomic region where they align. Single maps, however, have experimental errors comprising false extra cuts, false missing cuts, sizing issues, which are modeled with different error models (Sarkar 2006; Teague et al. 2010). In the second step, these map clusters are assembled by Gentig, a Bayesian maximum-likelihood assembler to generate consensus restriction maps (Anantharaman et al. 1999). This step addresses the experimental noise from individual maps. Finally, consensus maps are aligned back to the *in silico* reference map, which

helps us identify differences in our samples. In first iteration, the *in silico* maps are used as reference. In following iterations, the consensus maps generated in previous iteration are used as reference. Based on previous work, we determined that 8 iterations were sufficient to provide good accuracy and coverage for the human genome (Teague et al. 2010). Using this approach, we generated genome-wide physical restriction maps for the three genomes being analyzed. This assembly approach also helped us assemble complex genomic rearrangements into consensus maps.

2.3.5. Automated variation calling

Finally, consensus optical maps were aligned back to the *in silico* reference maps and differences between the two were tabulated under 5 categories: *i*) Insertions at locations where consensus map size exceeded reference map size; *ii*) Deletions at locations where consensus map size was less than reference map size; *iii*) Extra Cuts where consensus map presented a cut site while the reference did not; *iv*) Missing Cuts where consensus map was missing a cut site while the reference had one; and *v*) Other where the genomic rearrangement could not be explained with a single event. Following empirically determined filters were used to address errors from the system:

1) We discarded all insertions and deletions where the reference and consensus fragments were <5 kb in size. Additionally, the flanking reference and consensus fragments were required to be >4 kb in size.
2) We discarded all insertions and deletions with absolute size difference of less than 3 kb or where the absolute difference in size was less than 10% of the reference fragment size.

3) We discarded all events which had less than 10 supporting single molecule Rmaps. Based on empirical evidence, we believe this is a reasonable cutoff considering the high depth of coverage for our samples.
4) We discarded extra and missing cuts if they occurred in, or were flanked by fragments less than that 4 kb in size.

These filters eliminated many errors associated with small fragment desorption and other sizing issues from Optical Mapping. While they remove some true calls, they greatly reduce the number of false positives from automated variation calling process. Finally, all the variants were subject to manual curation to remove other erroneous calls.

2.3.6. Optical Map Coverage Analysis

The HMM coverage analysis algorithm summarizes the optical map alignments by a single number (midpoint) that represents locations, which are then modeled as realizations of a non-homogeneous Poisson process. The non-homogeneity of alignment data, which arises due to varied cognate site density across the genome (for restriction endonuclease), is accounted for by using alignment data from a normal genome, which are used to define intervals with counts that follow a negative binomial distribution. These counts are then modeled by a Hidden Markov Model, which incorporates spatial dependence in the data and allows more natural estimation of certain parameters. As a result, sample genome is partitioned into high, normal and low copy number states (Sarkar 2006; Sarkar et al. 2012).

2.4. Results and Discussion

2.4.1. Raw mapset collection

We collected Rmaps, as described in the previous section. We used BamHI restriction endonuclease (cognate site 5'-G|GATCC-3') for this work because of two reasons. First, BamHI yields an average fragment size of 9.18 kb for the BamHI restriction maps generated *in silico* from human genome reference sequence (build 37). This average fragment size and size distribution of underlying fragments correspond well with our current microscopy and machine vision capabilities. Second, BamHI is methylation insensitive and thus, not affected by methylation status of human genomic DNA. We let the data collection proceed until raw datasets, containing Rmaps greater than 300 kb, were approximately 300-fold deep in coverage of the *in silico* reference maps for each of the three samples. For example, if we approximate human reference sequence to be 3 gigabases long, we collected 900 gigabases worth of Optical Mapping data for each sample, which reflects 300 fold depth of coverage. One of the reasons for collecting deep datasets is that single Rmaps contain experimental errors like false cuts, missing cuts and sizing errors etc. By collecting multiple Rmaps that represent each locus across the genome, errors associated with individual Rmaps are corrected by assembly processes. The final datasets contain close to 2 million Rmaps for each of the analyzed samples. The Rmaps range in size from 300 kb to 2,500 kb and average close to 400 kb. Overall, these raw mapsets reflect 278, 270 and 325 fold depth of coverage for Normal, MM-S and MM-R samples, respectively. Table 2-1 details key statistics associated with Rmaps collected from these samples.

2.4.2. Iterative assembly

After data collection was complete, the Rmaps maps were fed to an iterative assembler (Valouev, Li, et al. 2006; Valouev, Zhang, et al. 2006; Valouev, Schwartz, et al. 2006; Teague et

al. 2010). Based on the information content of individual Rmaps (order and size of fragments), they were first aligned and placed on the *in silico* reference maps. Accordingly, Normal, MM-S and MM-R aligned datasets contain 752,218, 566,463 and 479,959 Rmaps in final assemblies, corresponding to 105, 80 and 70 fold coverage of the human reference genome, respectively (Fig. 2-2a,b,c). The alignment process clustered the Rmaps, which were then assembled into genome-wide contigs. The final consensus maps, generated after iterative assembly, spanned more than 99% of the reference human genome for all three samples and provided a scaffold for comprehensive discernment of structural variation. Table 2-2 provides a summary of alignment and assembly statistics. Table 2-3 lists final depth of coverage, by chromosome, for each assembly. One can clearly observe low coverage for chromosome 13 in MM-S and MM-R samples, relative to overall depth of coverage for these genomes. The coverage, approximately half of the expected for a normal karyotype, reflects loss of one copy of chr13 (monosomy 13) in MM-S and MM-R samples. Monosomy 13 is a common genomic aberration associated with multiple myeloma and is known to be associated with a poor prognosis (Walker et al. 2010).

2.4.3. Structural variation analysis

Consensus optical maps, constructed from iterative assembly, generate a genome-wide scaffold providing nearly telomere-to-telomere coverage of the genome. We have developed an automated pipeline that compares this scaffold to *in silico* reference maps to identify differences between the two, termed Optical Structural Aberrations (OSAs). OSAs are classified into five categories: insertions, deletions, extra cuts, missing cuts and other complex rearrangements. Fig. 2-3 presents prototypic examples for each of these OSAs. After structural variation analysis, we identified close to 2,000 OSAs in each of the tumor samples (Table 2-4,

Fig. 2-4). We identified 139, 149 and 176 deletions in Normal, MM-S and MM-R samples, respectively, ranging in size from ~3 kb to 180 kb with mean size ~10 kb and median size ~6 kb (Table 2-5). A histogram of deletion size distribution in these samples is presented in fig. 2-5. Similarly, we identified 450, 428 and 384 insertions in Normal, MM-S and MM-R samples, respectively, ranging in size from ~3 kb to 367.5 kb with mean size ~10 kb and median size ~5 kb (Table 2-6). A histogram of insertion size distribution in these samples is presented in fig. 2-6. More importantly, we characterized close to 100 complex OSAs in each of these samples; these variant structures cannot be explained by a single alteration and reflect more profound rearrangement of the genome at these loci. There are two important points that I want to highlight here. First, these OSAs include both, germline polymorphisms and somatic structural variants. In chapter 3, using a combination of Optical Mapping and DNA sequencing approaches, I will distinguish somatic structural variants from germline polymorphisms. Second, it might seem surprising that the total number of OSAs is approximately the same in normal and tumor (MM-S and MM-R) samples. Even more surprisingly, for some categories like insertions, the number of OSAs is lower in tumor samples when compared to the normal sample. This can be explained by sample level differences in data quality and subsequent filtering of variant calls based on empirically defined filters for variation calling.

2.4.4. Copy number variation analysis

The alignment of Rmaps to *in silico* human reference maps can serve as an indicator of copy number. In general, a smaller number of Rmaps will be expected to align to the reference in regions with deletions. Conversely, a larger number of maps will be expected to align to the reference in regions with amplifications. This expectation is analogous to copy number analysis

using DNA sequencing read data, where a deviation from average in the number of sequencing reads aligning to any genomic location serves to indicate copy number changes (Alkan et al. 2011). However, because of local differences in restriction site densities along a single chromosome or across chromosomes, differences are observed in depth of Rmap coverage even across a normal genome. This is clearly highlighted in Normal sample, where the depth of coverage ranges from 80 fold to 118 fold across chromosomes (Table 2-3). To account for these local differences, the depth of coverage from a test sample needs to be normalized against a reference sample. Based on these principles, previous members of our lab developed a computational approach that uses a Hidden Markov Model (HMM) based algorithm to identify somatic copy number changes in a test genome (Sarkar 2006; Sarkar et al. 2012; Ray et al. 2013). Accordingly, we have used the Normal sample as reference and compared the depth of coverage in MM-S and MM-R tumor samples with the depth of coverage in Normal sample to identify somatic copy number changes across these tumor genomes.

Upon HMM copy number analysis of MM-S and MM-R samples, these tumor genomes were partitioned into high, normal and low copy number states. In the MM-R genome, we observed 37 copy number breakpoints that were associated with genomic gains and losses. Surprisingly, these somatic copy number aberrations spanned approximately one-third of the total genome length and were generally associated with chromosomal ends (Fig. 2-7; red: copy number gain; blue: copy number loss; grey: reference gaps).

2.4.5. Identifying structural variation that underlies copy number variation

Copy number breakpoints are known to result from unbalanced genomic aberrations like unbalanced translocations, interstitial and terminal deletions etc. After identifying somatic

copy number breakpoints in the multiple myeloma tumor genomes, I sought to identify the structural rearrangements that resulted in these copy number changes. To this end, I visually inspected optical map assemblies at some of these breakpoints and observed a few consensus maps that aligned only partly to the reference at these locations. The remaining part of these consensus maps did not align to the genome at that location indicating that the segment originated from somewhere else in the genome (Fig. 2-8). Such consensus maps are expected because in later iterations during the assembly process, the consensus maps grow in a reference-independent manner. These chimaeric consensus maps were extracted and aligned back to whole genome *in silico* reference maps to identify the location/origin of the unaligned segment.

However, for some breakpoints, such chimaeric consensus maps were not observed. We attribute this to local complexities in assembly process or an over-abundance of reference-like maps in these locations, which prevents the assembly of consensus maps reflective of genomic rearrangements. To identify genomic rearrangements in these breakpoints, we designed a manual assembly approach (Fig. 2-9). This approach can be broken down into four basic steps: *i*) Make a list of all the copy number breakpoints for which the underlying structural aberration is unknown; *ii*) For each breakpoint, extract a subset of *in silico* reference map that flanks the breakpoint. In our implementation, we extracted a 500 kb region upstream and downstream of the breakpoint, totaling to 1 Mb. Such small reference segments allow quicker assembly and also allow us to perform more iterations to assemble the genomic rearrangement at the breakpoint; *iii*) Perform iterative assembly to 8 or more iterations on each of these reference maps using input mapsets from MM-S and MM-R samples; and *iv*) Identify partly aligning,

chimaeric consensus maps from the final assemblies, align them back to whole genome *in silico* reference maps to identify the origin of unaligned segment of the consensus maps.

Using the information from optical map assemblies and from the manual approach, we were able to resolve the structural rearrangements that explained the origin of 24 out of 37 copy number breakpoints (Table 2-7). These aberrations include unbalanced translocations, interstitial deletions, chromosomal truncations, tandem duplications and more complex rearrangements. The copy number changes represent canonical and non-canonical multiple myeloma genomic losses and gains and examples of complex events, drawn from fig. 2-7, are described here along with underlying structural aberrations.

1p loss: We identified a 112 kb deletion at 1p32.3 involving the genes FAF1 and CDKN2C (Fig. 2-10). The other chr1 allele has a 72 Mb long deletion, del(1)(p34.1;p13.1), encompassing this locus (Fig. 2-7) and leading to homozygous loss of FAF1 and CDKN2C. The deletion of negative cell cycle regulator CDKN2C is known to be a key early aberration in multiple myeloma (Leone et al. 2008).

Translocation(11;14): The immunoglobulin heavy chain locus on chr14 (IgH@) undergoes gene rearrangement, sequential rounds of somatic hypermutation and class switch recombination during cell differentiation and maturation of plasma cells (Morgan et al. 2012). These DNA modification processes introduce mutations and double stranded DNA breaks, and sometimes lead to aberrant chromosomal translocations and activation of oncogenes. The t(11;14)(q13;q32) translocation involves the IGH@ locus at 14q32 and the cyclin CCND1 at 11q13, and leads to overexpression of CCND1. CCND1 gene encodes for cyclin D1 and plays an important role in cell cycle G1/S transition. It is suggested that dysregulation of cyclin D1

expression upon t(11;14)(q13;q32.33) translocation leads the plasma cells towards tumorigenesis (Janssen et al. 2000). This translocation is reported for ~15% - 20% of multiple myeloma patients (Fonseca et al. 2003; Bergsagel & Kuehl 2005) and is believed to be one of the primary events associated with myelomagenesis. The chr11 breakpoint varies from 1 kb to 1 Mb upstream of CCND1 (Walker et al. 2013). We identified t(11;14) in both, MM-S and MM-R samples (Fig. 2-11). Consistent with previous reports, the breakpoint on chr11 is ~2 kb upstream of CCND1. We also observed del(14)(14q32.33) following the translocation breakpoint on chr14 and amplification of CCND1 locus due to another translocation between chr11 and chr14 (Fig. 2-7).

Monosomy 13: From Optical Map coverage analysis, we identified loss of one copy of chr13. Monoallelic loss of 13q is one of the most frequent abnormalities (~50% patients) in multiple myeloma and is known to be associated with a poor prognosis (Walker et al. 2010).

17p loss: Deletions involving TP53 locus occur in ~10% of untreated multiple myeloma patients (Kuehl & Bergsagel 2012). We observed a ~12 Mb deletion in 17p region, which includes the gene TP53. In MM-S and MM-R samples, this deletion forms part of complex genomic architecture, where two genomic segments (2 Mb - 13.89 Mb and 15.89 Mb – 16.23 Mb) are deleted and the middle region (13.89 Mb – 15.89 Mb) is inserted in an inverted orientation (Fig. 2-12).

Inversions: We identified previously reported 1.1 Mb germline inversion polymorphism at 16p12.1 (Antonacci et al. 2010). Also, we identified a 4.71 Mb somatic inversion at 14q21.3-q22.1 locus in both, MM-S and MM-R samples (Fig. 2-13).

Other translocations/copy number changes: We identified a translocation, t(2;10)(q33.3;p12.2) that leads to loss of ~35 Mb at q-ter of chr2 and amplification of ~22 Mb at p-ter of chr10 (Fig. 2-14). Another translocation between chr17 and chr19, t(17;19)(q21.31;19q13.43), leads to amplification of ~37 Mb at q-ter of chr17 and loss of ~1Mb at q-ter of chr19 (Fig. 2-15). In another event, ~26 Mb region from q-ter of chr18 is amplified in a fusion at the end of chr9 (Fig. 2-16).

In summary, using Optical Mapping, we have identified and characterized a large number of somatic structural rearrangements that lead to large scale copy number changes in the multiple myeloma genome. Additionally, we have highlighted a few aberrations that are acquired later in pathogenesis, as they are observed only in the MM-R sample (also see table 2-7), and might affect molecular pathways leading to drug resistance in multiple myeloma.

2.5. Conclusions

In conclusion, we have demonstrated that Optical Mapping provides long-range structural information about the genome because of its ability to analyze long DNA molecules, ranging in size from 300 kilobases to megabases. Consequently, we have identified a large number of optical structural aberrations in the multiple myeloma genomes. Additionally, we have identified a large number of somatic copy number aberrations and uncovered the structural rearrangements that underlie these common and uncommon multiple myeloma related copy number changes. Finally, from our analysis of copy number changes, we have provided evidence for increasing mutational complexity with tumor progression, which might have implications in multiple myeloma pathogenesis and drug resistance mechanisms.

As has been recently shown, cancer genomes are mutated in profound ways and include single nucleotide variants, small indels, structural rearrangements and larger copy number changes. In chapter 1, I reviewed the limitations of currently used genome analysis systems in identifying structural variation. In this chapter, we have shown that Optical Mapping is capable of overcoming many of those limitations in identifying genomic structure and consequently, copy number and structural variation. It is a whole genome approach with high resolution (3 kb) and the ability to identify both, structural and copy number changes. The large size of DNA analytes (300 kb to megabases) allows Optical Mapping to unambiguously resolve structural variants that overlap repeat rich regions of the genome.

However, Optical Mapping has certain limitations. First, its resolution, in the current state, is limited to 3 kb. Although we identify extra cuts and missing cuts that might originate from single base differences, we cannot identify most of the variation less than 3 kb in size. Second, the current optical map assembly algorithms consider a single representation of the human genome and consequently, many hemizygously presenting variants are not identified. Finally, Optical Mapping generates approximate breakpoints for structural variants and copy number changes. To design downstream genotyping assays and functional experiments, it is important to know precise basepair localization for these variants.

To address these issues with Optical Mapping, we used paired-end DNA sequencing, which is ideally suited to discover small variants. In Chapter 3, I will discuss our use of DNA sequencing to identify somatic variants less than 3 kb in size. Additionally, I will present a comparison of Optical Mapping and DNA sequencing in their ability to identify structural variation and copy number changes. This comparison also allowed us to resolve many structural

variants and copy number changes to basepair resolution. Finally, I will illustrate how an integrative approach using Optical Mapping and DNA sequences allowed us to understand complex genomic rearrangements in this multiple myeloma genome.

2.6. Bibliography

Alkan, C., Coe, B.P. & Eichler, E.E., 2011. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5), pp.363–76.

Anantharaman, T., Mishra, B. & Schwartz, D., 1999. Genomics via optical mapping. III: Contiging genomic DNA. *International Conference on Intelligent Systems for Molecular Biology*, pp.18–27.

Antonacci, F. et al., 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature genetics*, 42(9), pp.745–50.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–9.

Bergsagel, P.L. & Kuehl, W.M., 2005. Molecular pathogenesis and a consequent classification of multiple myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(26), pp.6333–8.

Boeva, V. et al., 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(3), pp.423–5.

Cai, W. et al., 1998. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 95(7), pp.3390–5.

Cai, W. et al., 1995. 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), pp.5164–8.

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

Fonseca, R. et al., 2003. The recurrent IgH translocations are highly associated with nonhyperdiploid variant multiple myeloma. *Blood*, 102(7), pp.2562–7.

Janssen, J.W.G. et al., 2000. Concurrent activation of a novel putative transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). *Blood*, 95(8), pp.2691–2698.

Kandoth, C. et al., 2013. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471), pp.333–9.

Kim, J. et al., 2012. Macrophages and mesenchymal stromal cells support survival and proliferation of multiple myeloma cells. *British journal of haematology*, 158(3), pp.336–46.

Korbel, J.O. et al., 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849), pp.420–6.

Kuehl, W.M. & Bergsagel, P.L., 2012. Molecular pathogenesis of multiple myeloma and its premalignant precursor. *The Journal of clinical investigation*, 122(10), pp.3456–63.

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

Leone, P.E. et al., 2008. Deletions of CDKN2C in multiple myeloma: biological and clinical implications. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(19), pp.6033–41.

Ley, T.J. et al., 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218), pp.66–72.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80.

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

Mitelman, F., Johansson, B. & Mertens, F., 2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer*, 7(4), pp.233–45.

Morgan, G.J., Walker, B. a & Davies, F.E., 2012. The genetic architecture of multiple myeloma. *Nature reviews. Cancer*, 12(5), pp.335–48.

Ray, M. et al., 2013. Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis. *BMC genomics*, 14, p.505.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–52.

Sarkar, D., 2006. On the analysis of optical mapping data. Ph.D. thesis, University of Wisconsin, Madison.

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

Schwartz, D.C. et al., 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (New York, N.Y.)*, 262(5130), pp.110–4.

Stephens, P.J. et al., 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276), pp.1005–10.

Sudmant, P.H. et al., 2010. Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, 330(6004), pp.641–6.

Teague, B. et al., 2010. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24), pp.10848–53.

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

Valouev, A., Schwartz, D.C., et al., 2006. 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), pp.15770–5.

Valouev, A., Zhang, Y., et al., 2006. Refinement of optical map assemblies. *Bioinformatics (Oxford, England)*, 22(10), pp.1217–24.

Walker, B. a et al., 2006. Integration of global SNP-based mapping and expression arrays reveals key regions, mechanisms, and genes important in the pathogenesis of multiple myeloma. *Blood*, 108(5), pp.1733–43.

Walker, B. a et al., 2010. A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, 116(15), pp.e56–65.

Walker, B.A. et al., 2013. Characterization of IGH locus breakpoints in multiple myeloma indicates a subset of translocations appear to occur in pregerminal center B cells. *Blood*, 121(17), pp.3413–9.

Weinstein, J.N. et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–20.

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

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

2.7. Figures

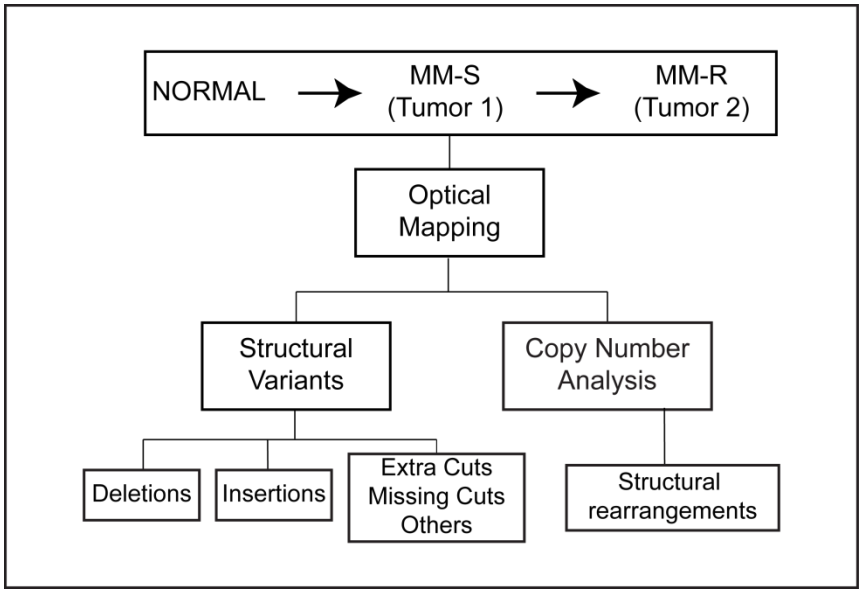

Fig. 2-1: Study design for multiple myeloma genome analysis using Optical Mapping.

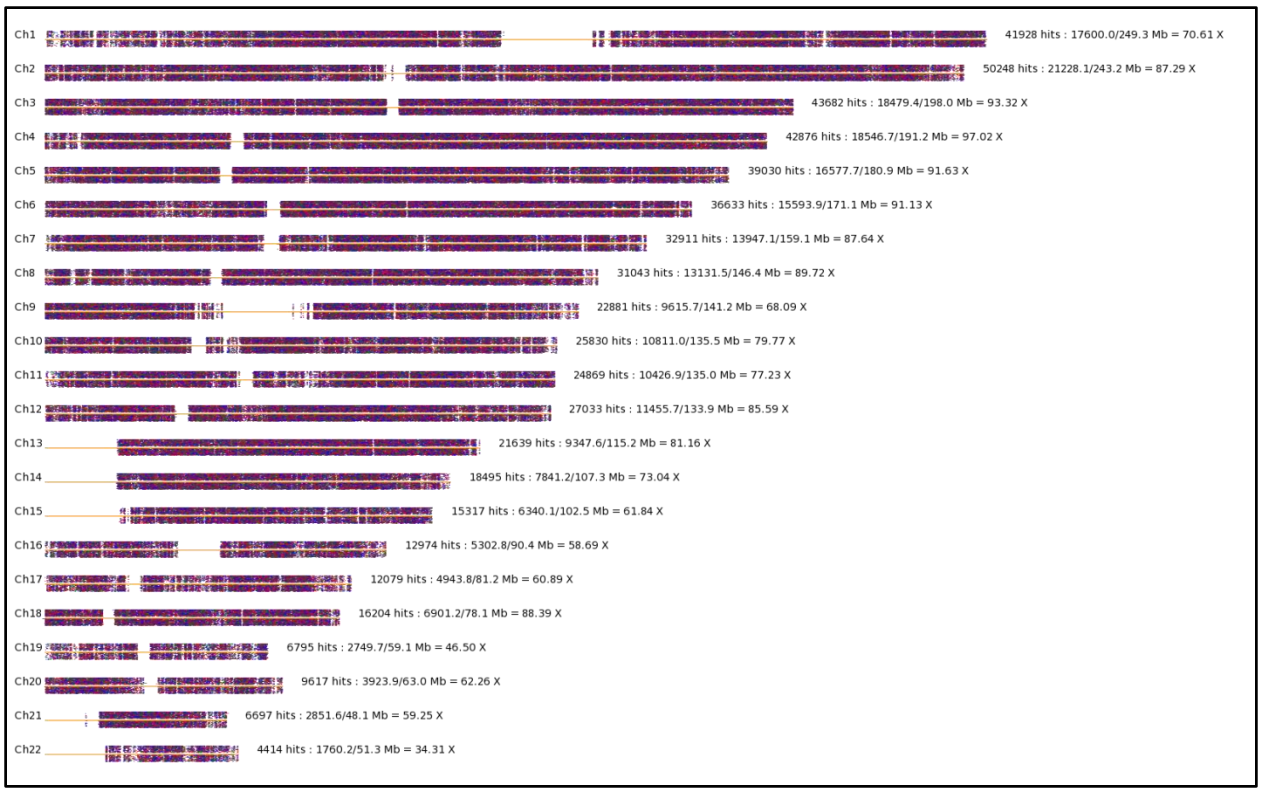

Fig. 2-2a: Aligned Rmaps from Normal sample. The numbers after each chromosome represent total number of Rmaps that align to each chromosome and depth of coverage based on initial alignments.

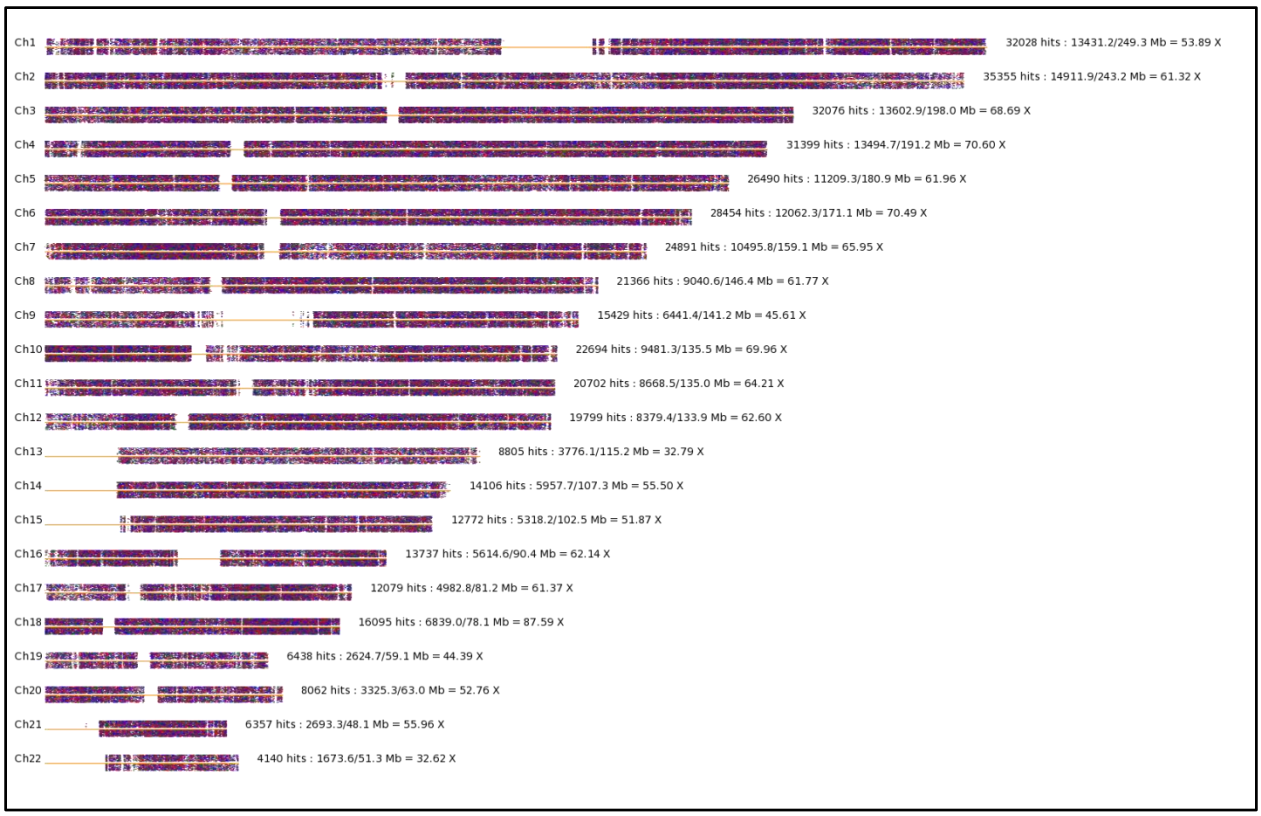

Fig. 2-2b: Aligned Rmaps from MM-S sample. The numbers after each chromosome represent total number of Rmaps that align to each chromosome and depth of coverage based on initial alignments.

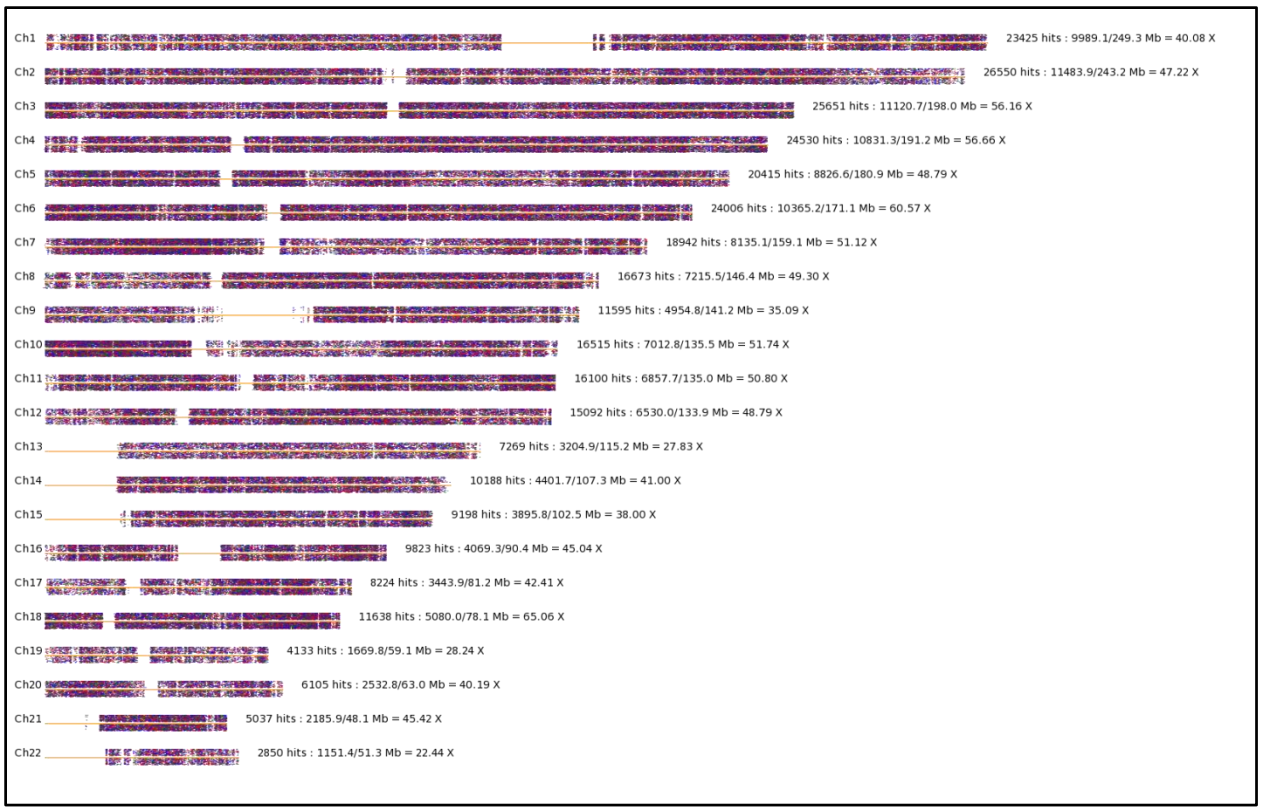

Fig. 2-2c: Aligned Rmaps from MM-R sample. The numbers after each chromosome represent total number of Rmaps that align to each chromosome and depth of coverage based on initial alignments.

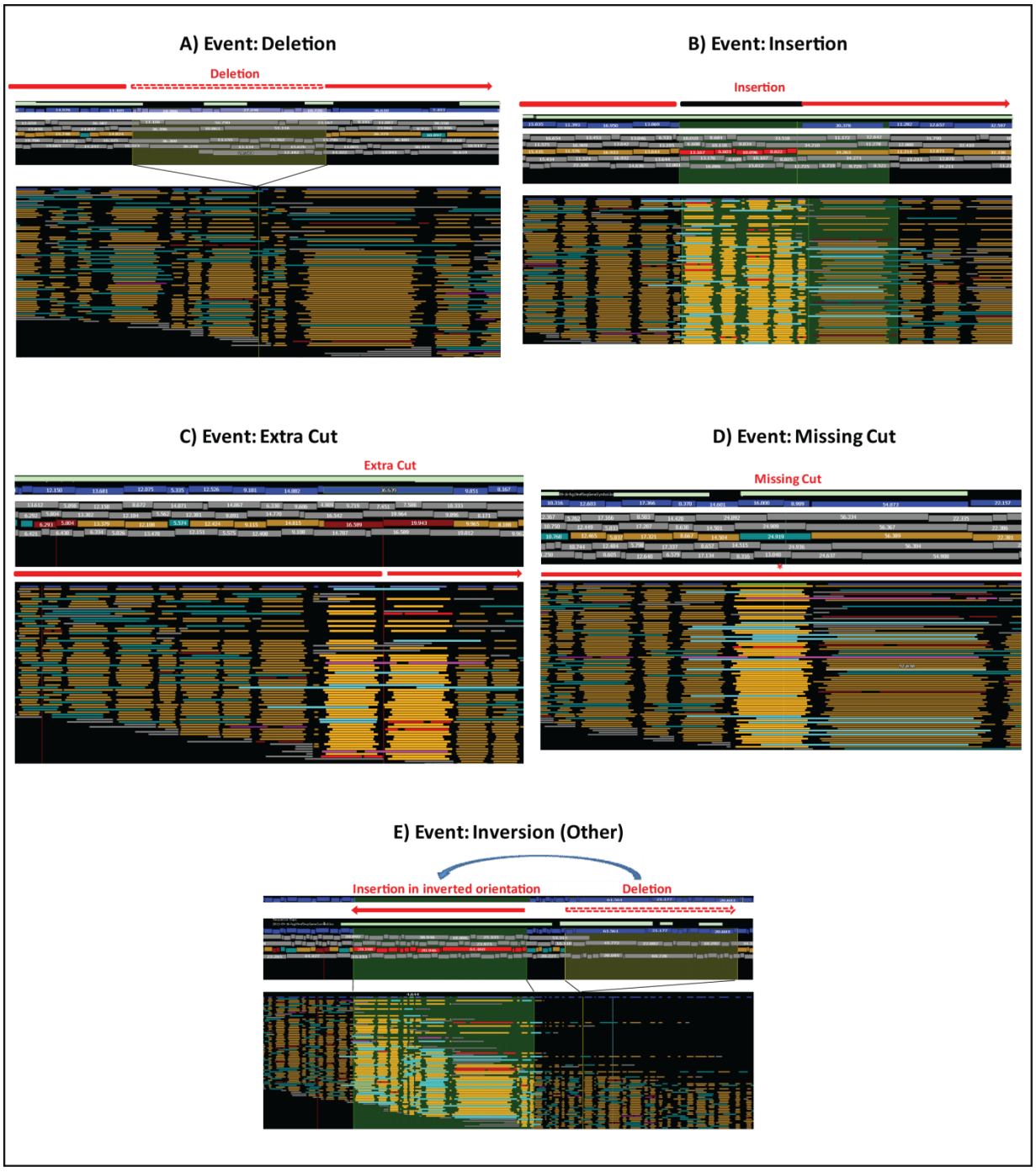

Fig. 2-3: Examples of optical structural aberrations identified from Optical Mapping automated variation calling pipeline. a) Insertions; b) Deletions; c) Extra cuts; d) Missing cuts; and e) Complex variants are shown. In each panel, the image on top shows alignments of consensus maps to *in silico* generated reference maps and the image at bottom shows Rmaps that were assembled to generate the consensus map.

Fig. 2-4: Optical structural aberrations identified in Normal, MM-S and MM-R samples. Chromosomes are represented by golden traces; sequence gaps by grey traces; structural aberrations in Normal sample by green markers, structural aberrations in MM-S sample by red markers; and structural aberrations in MM-R sample by blue markers.

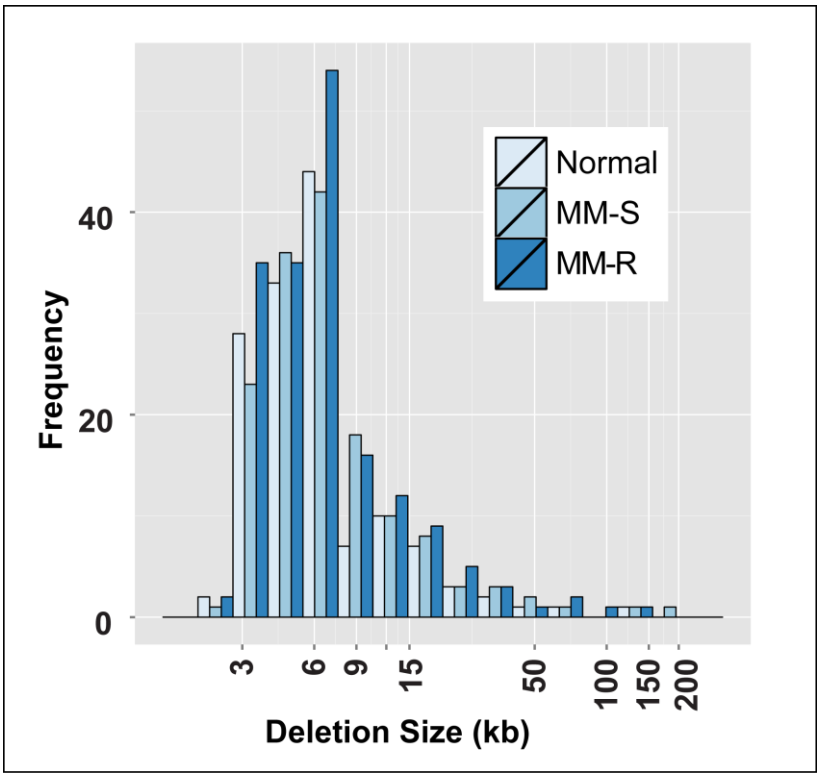

Fig. 2-5: Distribution of deletion sizes in Normal, MM-S and MM-R samples. The size of deletions (in kilobases) is plotted along X-Axis and the frequency of deletions is plotted along Y-Axis.

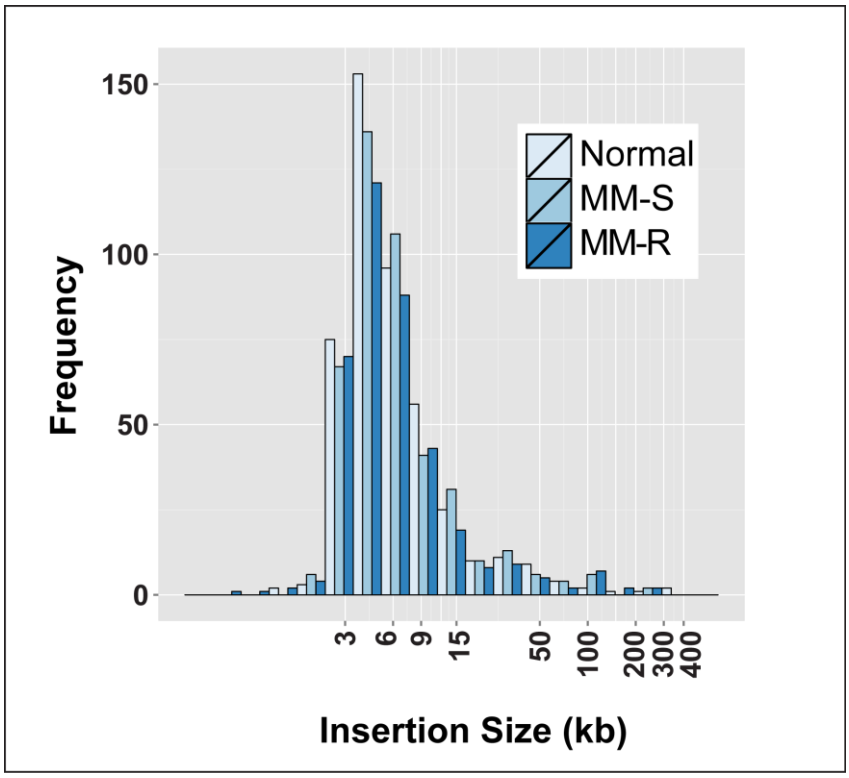

Fig. 2-6: Distribution of insertion sizes in Normal, MM-S and MM-R samples. The size of insertions (in kilobases) is plotted along X-Axis and the frequency of insertions is plotted along Y-Axis.

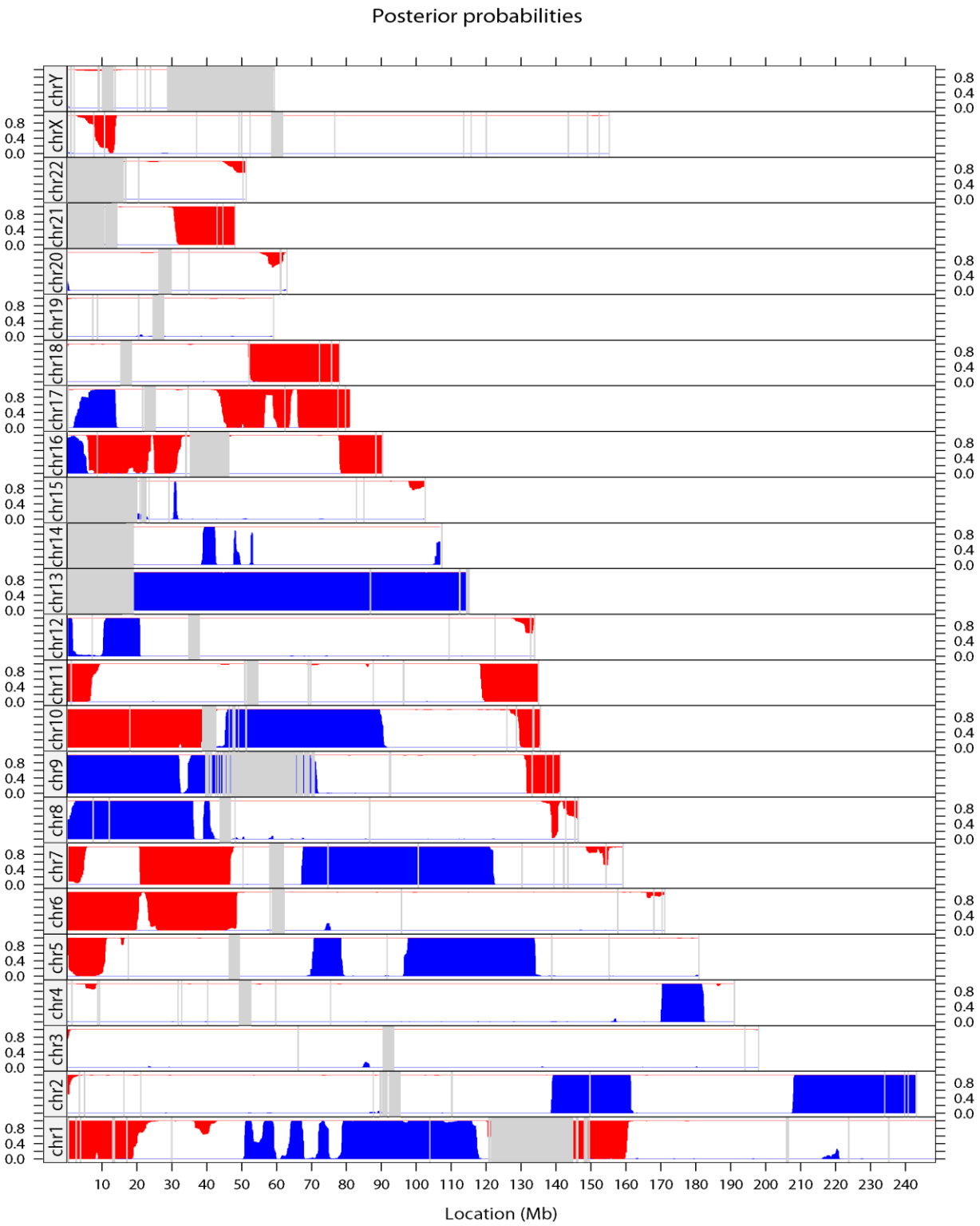

Fig. 2-7: Copy number changes in MM-R sample identified using HMM analysis. Blue color highlights reduced copy number regions; red color highlights increased copy number regions; and light grey color highlights gaps in the human reference sequence.

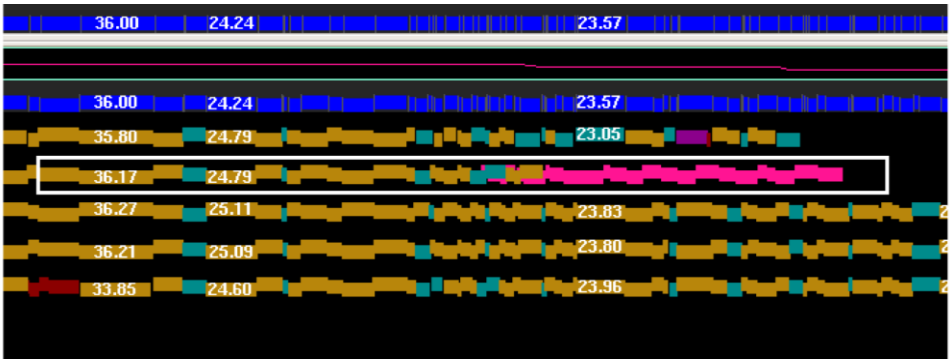

Fig. 2-8: A prototypical chimaeric consensus map that aligns partly to the reference map. The trace in blue at the top shows *in silico* generated human reference map. The numbers are fragment sizes in kilobases. The traces below the blue trace (golden and other colours) show consensus maps generated from optical map assembly for this region, aligned back to the reference maps. While most of the consensus maps align well (many golden fragments), one of the consensus maps (in box with white ouline) aligns partly to the reference. The pink overhang does not align to the reference at this location, indicating this consenus map is chimaeric.

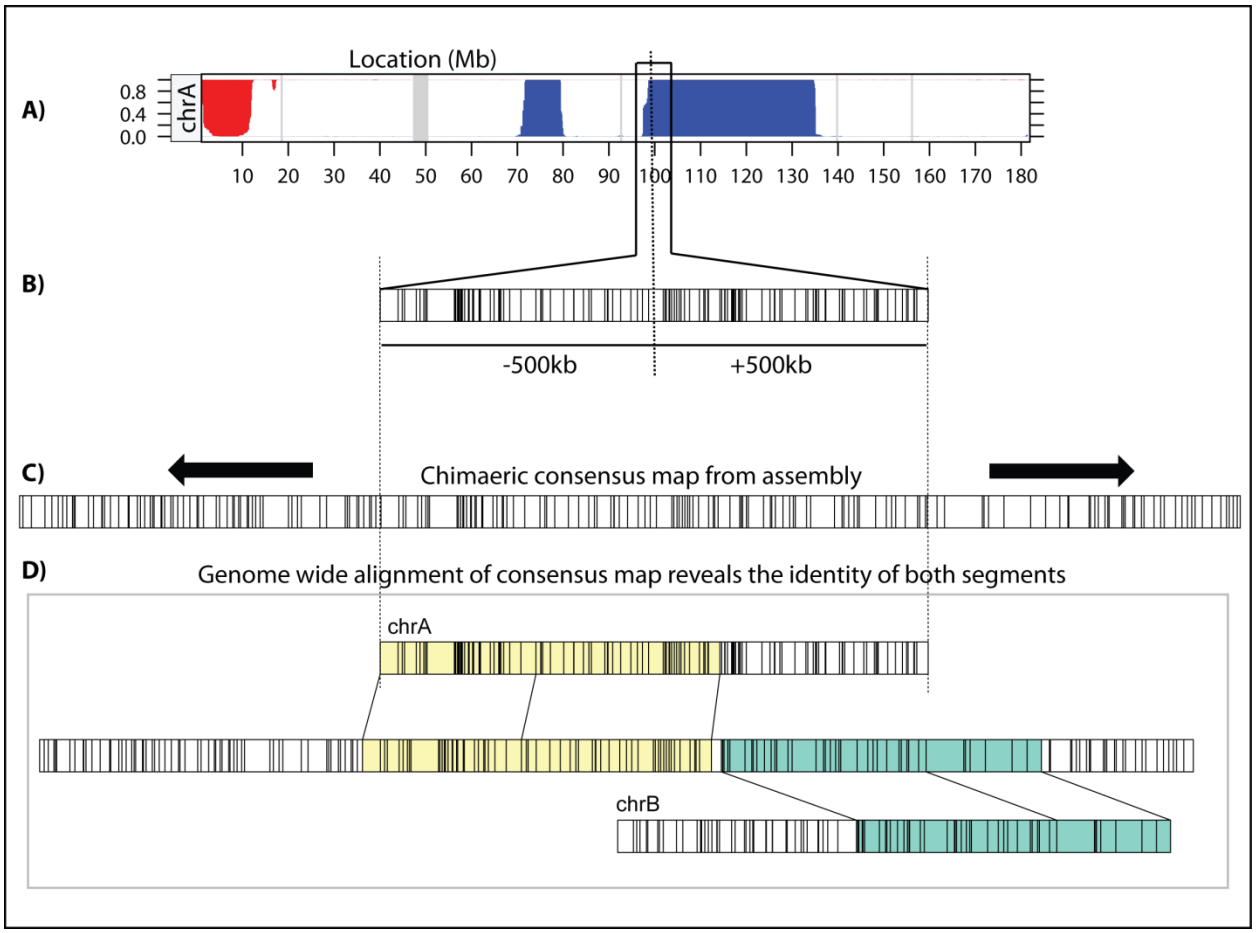

Fig. 2-9: An approach for optical map assembly of structural rearrangements underlying copy number changes. For any copy number breakpoint where the underlying genomic rearrangement is not known (A), a megabase long stretch of *in silico* reference maps, with 500 kilobases on either side of the breakpoint, is extracted (B). This *in silico* reference map is submitted to iterative assembly for growth using Rmaps collected from this sample (C). The assembly process, sometimes, generates a chimaeric consensus map. Upon aligning this chimaeric consensus map to the entire set of human reference *in silico* maps, the genomic origin of both segments is identified (here chrA and chrB), thereby revealing the structural rearrangement underlying the copy number breakpoint (D).

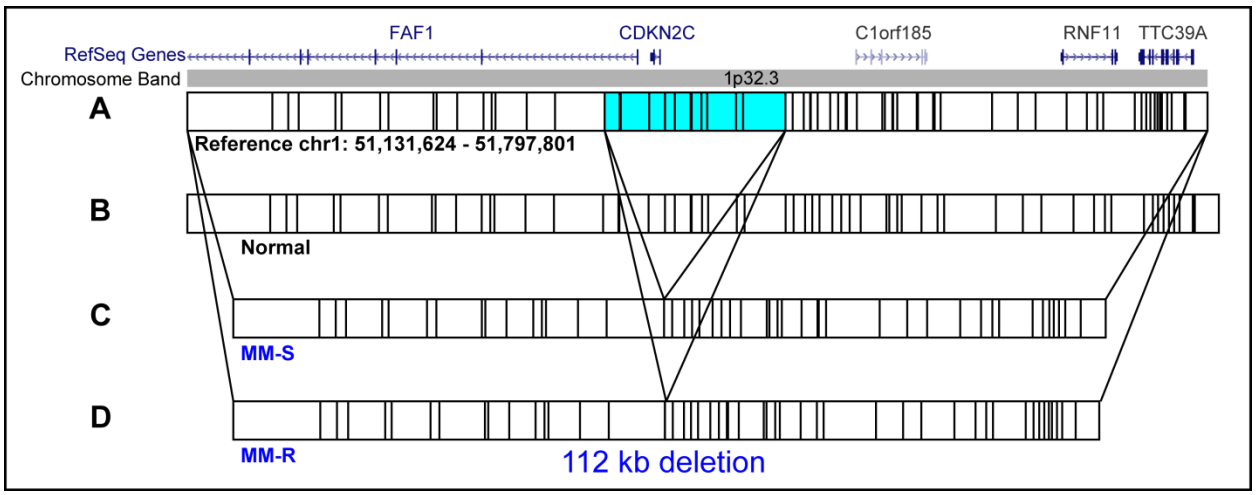

Fig. 2-10: CDKN2C/FAF1 deletion in multiple myeloma. The first track (A) shows *in silico* reference map for genomic locus 1p32.3 and is annotated with chromosomal band and RefSeq genes (tracks above A) from UCSC genome browser. The alignments of consensus optical maps from Normal (B), MM-S (C) and MM-R (D) assemblies to *in silico* reference map are shown. Normal sample matches the reference, whereas MM-S and MM-R consensus maps show ~112 kb deletion (chr1: 51.388 Mb – 51.500 Mb; highlighted in light blue color in reference) that overlaps the genes FAF1 and CDKN2C.

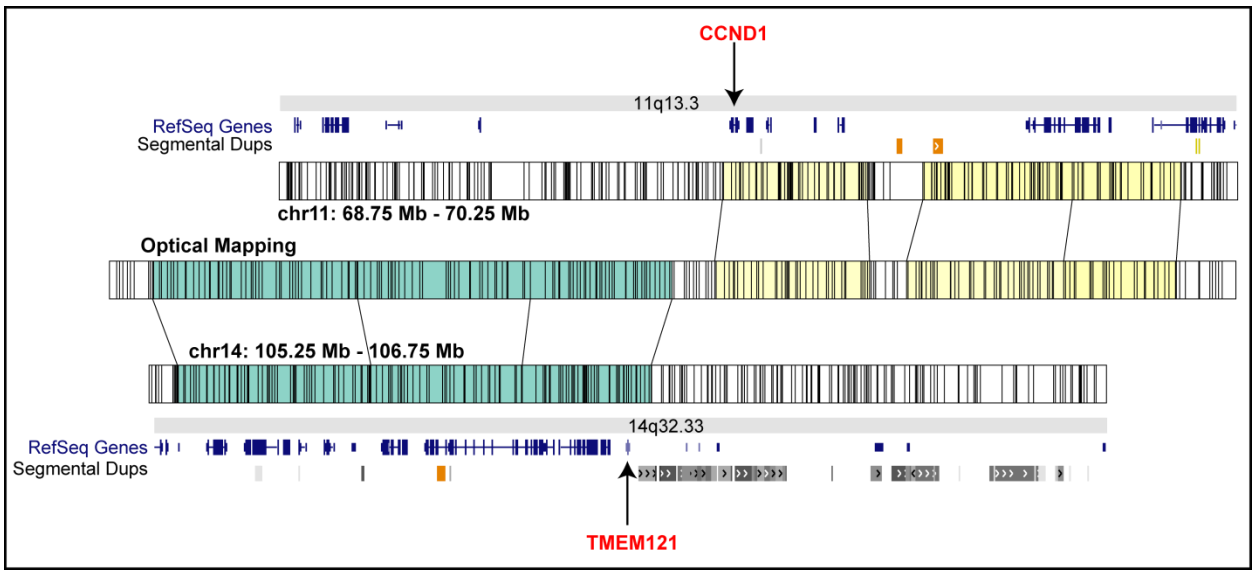

Fig. 2-11: t(11;14)(q13.3;q32.33) IGH@-CCND1 translocation in multiple myeloma. The top track (marked chr11: 68.75 Mb - 70.25 Mb) and the bottom track (marked chr14: 105.25 Mb - 106.75 Mb) show *in silico* reference maps for these genomic loci and have been annotated with chromosomal bands, RefSeq genes and Segmental Duplications (top to bottom) from UCSC genome browser. The track in the middle (Optical Mapping) shows consensus optical map from iterative assembly. The lines connecting these tracks show alignment of the consensus optical map to *in silico* reference maps and indicate that the consensus map aligns partly to chr11 and partly to chr14, thereby revealing a translocation. The chr11 breakpoint is upstream of gene CCND1 and chr14 breakpoint lies in immunoglobulin heavy chain locus.

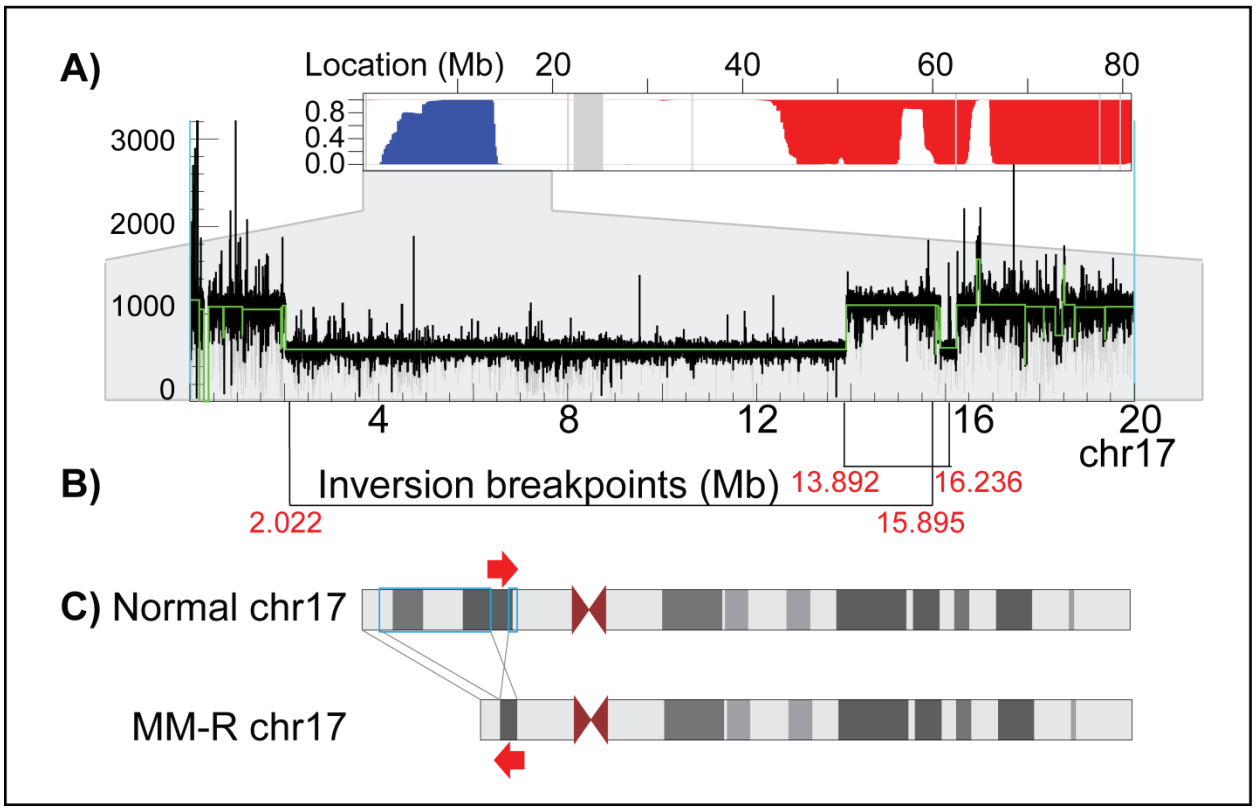

Fig. 2-12: Two deletions and an associated inversion explain copy number (CN) changes at p-ter of chr17 and reveal complex genome architecture. A) Posterior probability plot of CN state in chr17 from MM-R *vs.* Normal HMM CN analysis of Rmap alignments. For each bin, length of blue bar (bottom to top) or red bar (top to bottom) indicates the probability of low or high CN state, respectively. Grey regions highlight gaps in reference sequence; white regions show normal CN regions. A low CN state is observed from ~2 Mb-13.89 Mb. Further CN analysis of chr17: 1-20 Mb using CNVnator presents two regions of reduced CN: 2.022 – 13.892 Mb and 15.895 – 16.236 Mb (Y-Axis represents number of aligned reads in 1000 bp window). B) Optical Map assembly and DNA sequencing structural variation analysis reveal two inversion breakpoints (black lines; detailed in C) in this region. C) The segments 2.022 - 13.892 Mb and 15.895 - 16.236 Mb (regions with blue outline in Normal) are deleted and 2 Mb long intervening region (13.892-15.895 Mb, red arrows) is inserted in inverted orientation.

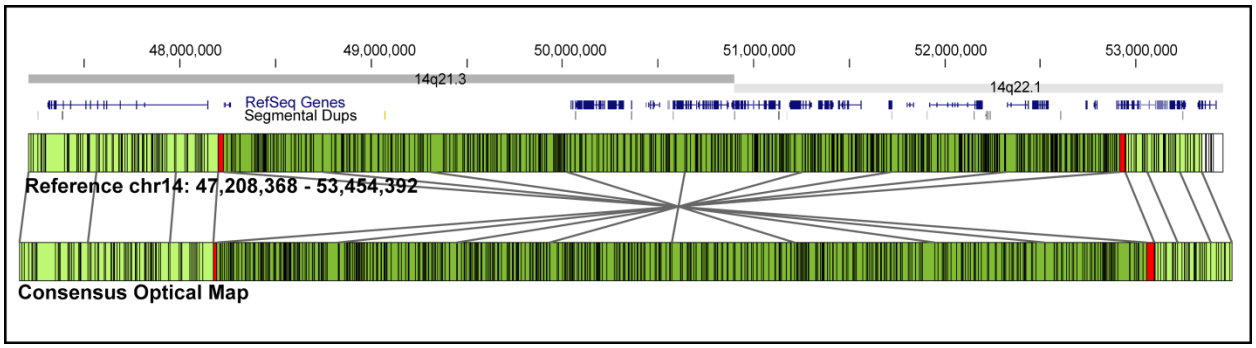

Fig. 2-13: A 4.71 Mb inversion on chr14 in MM-S and MM-R samples. The top track (labeled Reference chr14) represents *in silico* reference map for a 6.25 Mb region on chr14, and is annotated with chromosomal position, chromosomal bands, RefSeq genes and segmental duplications from UCSC genome browser. The bottom track represents consensus map for this region from Rmap iterative assembly. The alignment between reference and consensus maps is shown by connecting black lines. The alignment indicates inversion of a 4.71 Mb region; the fragments that contain the breakpoints are highlighted in red. The left breakpoint lies between chr14: 48,199,703 - 48,227,801 and the right breakpoint lies between chr14: 52,914,995-52,946,843.

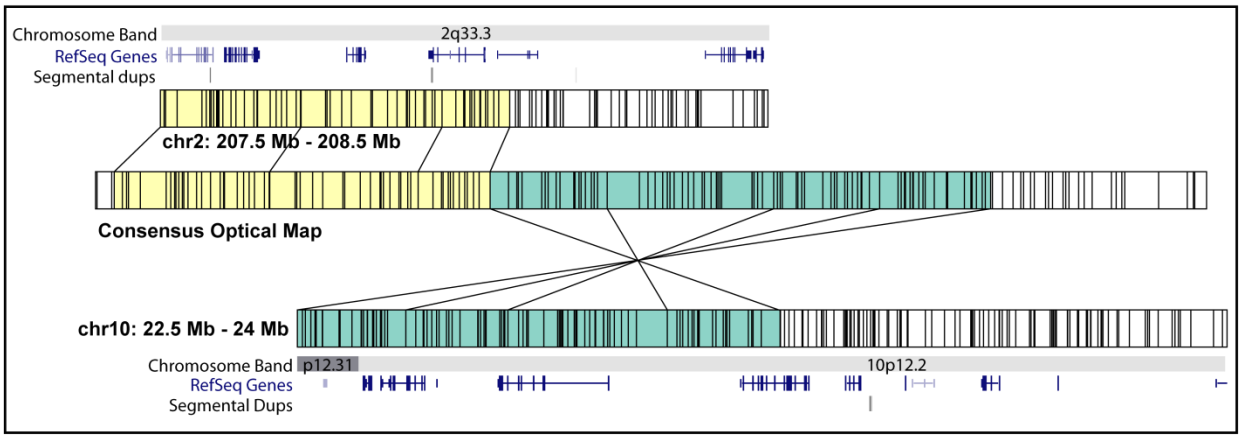

Fig. 2-14: A translocation between chr2 and chr10, t(2;10)(q33.3;p12.2), explains copy number loss on chr2 and copy number gain on chr10. The top track (marked chr2: 207.5 Mb – 208.5 Mb) and the bottom track (marked chr10: 22.5 Mb – 24 Mb) represent *in silico* reference maps for these genomic loci and have been annotated with chromosomal bands, RefSeq genes and segmental duplications (top to bottom) from UCSC genome browser. The track in the middle (Consensus Optical Map) shows consensus optical map from Rmap iterative assembly. The black lines connecting these tracks show alignment of the consensus optical map to *in silico* reference maps and indicate that the consensus map aligns partly to chr2 and partly to chr10, thereby revealing a translocation. The region following the breakpoint is lost on chr2 (chr2: 208.07 Mb - End), thereby showing a copy number of 1 (see Fig. 2). The region preceding the breakpoint is amplified on chr10 (23.27 Mb –Start), thereby showing a copy number of 3 (Also see fig. 2-7).

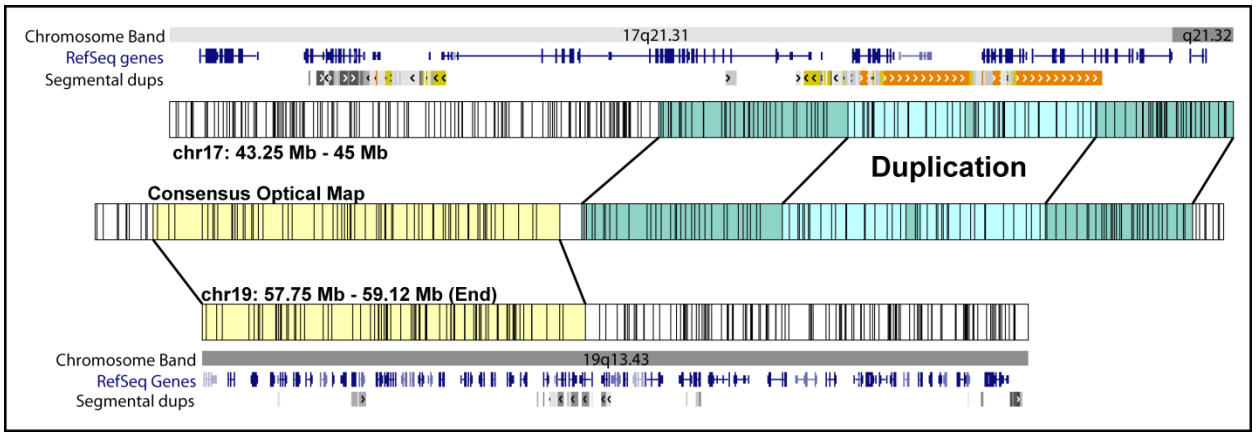

Fig. 2-15: A translocation between chr17 and chr19, t(17;19)(q21.31;q13.43), explains copy number gain on chr17 and copy number loss on chr19. The top track (marked chr17: 43.25 Mb – 45 Mb) and the bottom track (marked chr19: 57.75 Mb – 59.12 Mb) represent *in silico* reference maps for these genomic loci and have been annotated with chromosomal bands, RefSeq genes and segmental duplications (top to bottom) from UCSC genome browser. The track in the middle (Consensus Optical Map) shows consensus optical map from Rmap iterative assembly. The black lines connecting these tracks show alignment of the consensus optical map to *in silico* reference maps and indicate that the consensus map aligns partly to chr17 and partly to chr19, thereby revealing a translocation. The region following the breakpoint is lost on chr19 (chr19: 58.41 Mb - End), thereby showing a copy number of 1 (see fig. 2-7). The region following the breakpoint is amplified on chr17 (44.04 Mb – End), thereby showing a copy number of 3 (see fig. 2-7).

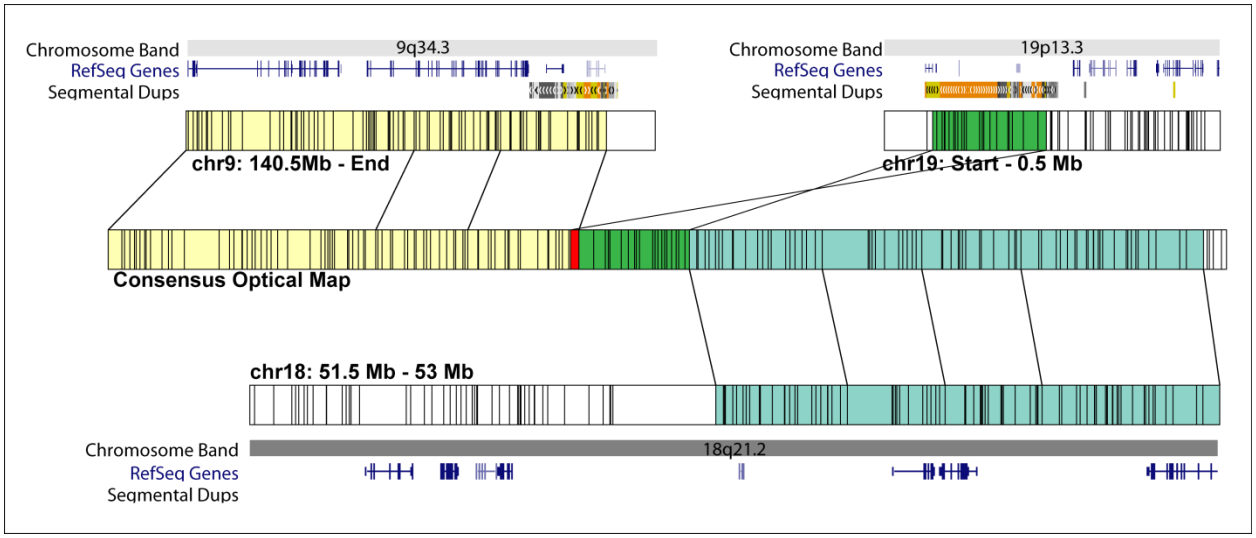

Fig. 2-16: A fusion at the end of chr9 explains copy number gain on chr18. The top tracks (marked chr9: 140.5 Mb –End and chr19: Start – 0.5 Mb) and the bottom track (marked chr18: 51.5 Mb – 53 Mb) represent *in silico* reference maps for these genomic loci and have been annotated with chromosomal bands, RefSeq genes and segmental duplications (top to bottom) from UCSC genome browser. The track in the middle (Consensus Optical Map) shows consensus optical map from Rmap iterative assembly. The black lines connecting these tracks show alignment of the consensus optical map to *in silico* reference maps and indicate that the consensus map aligns partly to chr9, chr19 and chr18.

We believe that the region aligning to chr19 is a reference assembly error and that this ~200 kb region is really a part of q-ter of chr9. We have based this on our analysis of Normal genome assemblies (Normal from this work and other assemblies from Teague et al. 2010).

The region following the breakpoint (blue color alignments) captures the fused segment from chr18 (chr18: 52.05 Mb – End), which fuses at the end of chr9 and thereby, shows a copy number of 3 (Also see fig. 2-7).

2.8. Tables

Table 2-1: Summary of Optical Mapping data collection for Normal, MM-S and MM-R samples.

Parameter	Normal	MM-S	MM-R
Number of molecules (#)	2,083,257	1,967,911	2,372,855
Total Mass (Gb)	832.945	807.297	980.905
Coverage (X)	~278	~270	~325
Average molecule size (kb)	399.83	410.23	413.39
Average fragment size (kb)	15.25	15.01	15.98
Fragments/molecule (#)	26.21	27.33	25.87
Days needed for collection	63	50	52
Surfaces used	450	290	319

Table 2-2: Summary of optical map alignment and assembly statistics for Normal, MM-S and MM-R samples.

Parameter	Normal	MM-S	MM-R
Input optical maps (#)	2,083,257	1,967,911	2,372,855
Input coverage (X)	~278	~270	~325
Assembled optical maps (#)	752,218	566,463	479,959
Assembled coverage (X)	~105	~80	~70
Consensus maps (#)	20,080	15,459	12,838
Average consensus map size (kb)	1,716.409	1,684.074	1,697.673
Sequence scaffold coverage (%)	99.65%	99.65%	99.5%

Table 2-3: Rmap depth of coverage per chromosome for Normal, MM-S and MM-R samples.

Chr	Normal	MM-S	MM-R
chr1	99.94	80.17	70.81
chr2	107.77	77.84	70.09
chr3	113.63	84.16	78.98
chr4	117.31	86.19	78.12
chr5	112.52	77.45	69.17
chr6	113.53	88.64	88.79
chr7	110.91	84.6	76.57
chr8	111.02	77.54	69.92
chr9	104.02	75.25	67.75
chr10	103.47	91.76	78.74
chr11	99.41	85.31	78.19
chr12	108.31	81.44	71.47
chr13	116.51	48.11	46.21
chr14	109.26	86.33	73.18
chr15	100.77	85.11	71.65
chr16	94.04	104.25	90.27
chr17	89.99	91.82	75.49
chr18	113.12	113.91	96.65
chr19	79.18	76.45	62.03
chr20	89.57	80.96	72.58
chr21	105.02	103.77	96.51
chr22	80.9	83.36	70.91
chrX	57.93	46.59	42.7
chrY	61.04	52.33	47.06
TOTAL	105	80	70

Table 2-4: Optical structural aberrations and the number of intersecting genes identified in Normal, MM-S and MM-R samples.

Type of event	**Normal**		**MM-S**		**MM-R**	
	Total	Number of genes	Total	Number of genes	Total	Number of genes
Insertions	450	234	428	213	384	190
Deletions	139	43	149	50	176	67
Extra Cuts	878	335	793	316	815	317
Missing Cuts	449	176	503	212	546	225
Other	88	58	96	60	115	70
Total	2004	846	1969	851	2036	869

Table 2-5: Size summary of deletions identified using Optical Mapping.

	Normal	**MM-S**	**MM-R**
Minimum (kb)	2.72	2.72	2.45
Mean (kb)	8.64	10.54	9.54
Median (kb)	5.78	5.99	5.92
Maximum (kb)	118.20	179.50	114.20

Table 2-6: Size summary of insertions identified using Optical Mapping.

	Normal	**MM-S**	**MM-R**
Minimum (kb)	1.17	1.66	0.62
Mean (kb)	10.87	10.46	11.19
Median (kb)	4.95	5.18	4.93
Maximum (kb)	367.53	274.67	282.50

Table 2-7: Copy number breakpoints identified in MM-R sample annotated with underlying structural rearrangements.

chr	start	end	Copy number State	Structure from Optical Mapping	Event description
chr1	656,307	19,193,483	HIGH	No	Unresolved, subclonal
chr1	50,907,765	117,406,209	LOW	Yes	Interstitial deletion
chr1	145,027,094	160,568,955	HIGH	No	Translocation (1;6) at right breakpoint
chr1	160,523,100	End		No	Loss of Heterozygosity
chr2	138,867,648	161,436,880	LOW	Yes	Interstitial deletion
chr2	208,007,180	242,954,292	LOW	Yes	Translocation (2;10) at left breakpoint
chr4	170,322,822	182,382,847	LOW	Yes	Interstitial deletion
chr5	28,898	10,963,144	HIGH	False Call	NA
chr5	70,678,962	78,556,479	LOW	Yes	Chromosomal truncation at left breakpoint
chr5	78,556,479	96,499,335	NORMAL	Yes	Tandem Duplication
chr5	96,499,335	134,206,720	LOW	Yes	Chromosomal truncation
chr5	134,206,720	180,750,424	NORMAL	No	Unresolved, Loss of Heterozygosity
chr5	180,039,800	End		Yes	Translocation(5;11) at left breakpoint
chr6	224,048	48,651,182	HIGH	No	Translocation(1;6) at right breakpoint
chr7	435,916	4,949,031	HIGH	False Call	NA
chr7	20,881,924	46,791,922	HIGH	Yes	Translocation t(7;16) at both breakpoints
chr7	67,225,303	122,071,003	LOW	Yes	Amplification, inversion and truncation at left breakpoint; translocation (7;12) at right breakpoint
chr8	146,436	36,187,296	LOW	No	Translocation (8;9) at right breakpoint
chr8	39,076,883	40,983,018	LOW	Yes	Deletion
chr8	138,753,274	140,674,356	HIGH	False Call	NA
chr9	120,912	32,225,396	LOW	No	Translocation(8;9) at right breakpoint
chr9	34,818,439	71,622,549	LOW	No	Unresolved, Right breakpoint is an approximation
chr9	131,601,057	141,107,212	HIGH	No	Unresolved
chr9	End	End		Yes	Translocation(9;18) at the end of chr9
chr10	200,298	42,879,738	HIGH		Split in next 3 events
chr10	0	4,887,601	HIGH	No	Unresolved
chr10	4,887,601	23,278,000	HIGH	Yes	Translocation t(2;10) at right breakpoint
chr10	23,278,000	39,154,900	HIGH	No	Unresolved
chr10	45,348,648	90,554,987	LOW	No	Unresolved
chr10	129,590,531	135,329,178	HIGH	False Call	NA
chr11	369,395	6,949,530	HIGH	False Call	NA

chr	start	end	Copy number State	Structure from Optical Mapping	Event description
chr11	118,638,018	134,832,095	HIGH	Yes	Translocation (5;11) at left breakpoint
chr12	298,612	20,981,741	LOW	Yes	Translocation(7;12) at right breakpoint
chr13	19,153,400	114,976,762	LOW	Yes	Monosomy chr13
chr14	38,929,283	42,600,725	LOW	One breakpoint	Translocation(14:21) at left, translocation(11;14) at right breakpoint
chr14	42,600,725	107,031,980	NORMAL	No	Loss of heterozygosity
chr15	30,653,006	31,407,295	LOW	Yes	Interstitial deletion
chr16	205,527	5,586,021	LOW	Yes	Translocation(7;16) at right breakpoint
chr16	6,139,046	31,693,579	HIGH	One breakpoint	Translocation(7;16) at left breakpoint
chr16	78,079,011	90,156,498	HIGH	Yes	Translocation t(7;16) at left breakpoint
chr17	3,757,149	14,040,686	LOW	Yes	Deletion, inversion
chr17	15,895,201	16,236,200	LOW	Yes	Deletion, inversion
chr17	44,146,692	81,068,128	HIGH	Yes	Translocation(17;19) at left breakpoint
chr18	52,474,910	77,859,828	HIGH	Yes	Translocation(9;18) at left breakpoint
chr21	30,643,661	48,200,250	HIGH	No	Translocation(14;21) at left breakpoint

Note: Events in yellow background represent copy number changes observed only in MM-R and not in MM-S tumor sample.

Chapter 3: Integrative approach using DNA Sequencing and Optical Mapping leads to comprehensive characterization of variation in multiple myeloma.

In this chapter, I will provide a brief introduction to the development of DNA sequencing methods. Then, I will describe how we have used paired-end DNA sequencing for an orthogonal analysis of multiple myeloma genomes studied in Chapter 2 using Optical Mapping. Finally, I will present and discuss our integrated findings from studying these genomes using Optical Mapping and DNA sequencing.

3.1. Introduction and motivation

3.1.1. Development of DNA sequencing methods

Sanger sequencing

The early methods for DNA sequencing were developed in mid to late 1970s by Sanger and colleagues (Sanger & Coulson 1975; Sanger et al. 1977) and Maxam & Gilbert (1977). Of these, chain-termination based Sanger sequencing was widely adopted because of its simplicity. In chain-terminator Sanger sequencing, primers are extended using dNTPs and terminated using ddNTPs along a single-stranded DNA template. The reaction mix contains template, primer, a DNA polymerase, three dNTPs (one labeled with ^{32}P) and a mix of deoxy and dideoxy NTPs for the fourth nucleotide. Primer extension is stochastically terminated by polymerase-mediated addition of a dideoxy nucleotide, resulting in a mixed population of extended primer sequences, or a sequence "ladder". The same termination process is then repeated for other nucleotides in separate reactions. The products are electrophoresed in parallel on a polyacrylamide gel, which resolves these bands to single base resolution. The pattern of bands is finally used to decipher DNA sequence for the template. Although revolutionary in that it

provided the first way to sequence DNA, such early approaches were very labor-intensive, low throughput and required careful disposal of radioisotopes.

A major advance that led to automation and use of Sanger sequencing in many early genome sequencing projects was the replacement of radioactive labels with fluorescent labels and consequent replacement of autoradiography detection methods with laser-induced fluorescence detection methods (Smith et al. 1986; Ansorge et al. 1986; Prober et al. 1987). This version was first developed by Leroy Hood's group at Caltech in collaboration with Applied Biosystems (ABI) in 1986 (Smith et al. 1986). In this method, the primers used for each of the four sequencing reactions were labeled with different fluorescent dyes. Upon template-defined polymerase extension of these primers in separate reactions, extension continues within a population of templates until stochastic termination happens upon incorporation of reaction-specific ddNTP. Reaction products from all four reactions were then pooled together and electrophoretically separated down a single polyacrylamide gel tube. Finally, fluorescence signals from electrophoretically separated products were detected near bottom of the gel tube by sequential excitation using laser lines, where the primer-defined fluorescence signal identified the incorporated ddNTP. This approach was later adapted for slab gel electrophoresis.

Although widely used, slab gel based sequencers were tedious to use because they required manual preparation of gels and loading of DNA samples. Additionally, their throughput was limited by the number of lanes per gel. The next major advance in DNA sequencing was marked by the development of capillary electrophoresis based sequencing methods in 1990s (Swerdlow & Gesteland 1990; Luckey et al. 1990; Swerdlow et al. 1991). In capillary sequencing

methods, the reaction products are automatically injected in capillaries filled with a flushable polymer sieving support, size separated and optically detected *via* fluorescence detection of resolved reaction products. Also, many capillaries (e.g. 96 or 384) can be run and analyzed in parallel, which increases the throughput many-fold. The development of capillary based sequencing systems (Dovichi 1997) enabled successful sequencing of the human genome (Lander et al. 2001; Venter et al. 2001). Our lab contributes to curation of the human genome build through comparisons with Optical Mapping findings (http://www.sanger.ac.uk/research/areas/bioinformatics/grc/).

Sanger sequencing methods have many advantages and disadvantages. Their main advantages are their long read lengths (up to 1000 bases) and high accuracy (99.997%). However, even in parallel capillary mode, they are limited by throughput, which is estimated at a maximum of around 1-2 megabases per day for commonly used 96-capillary instruments (Morozova & Marra 2008; www.appliedbiosystems.com). Additionally, they require a long and laborious process of clone library construction and sub-cloning for production of DNA sequencing templates. These steps are costly and omit genomic regions not amenable to cloning. As such, Sanger sequencing cannot support the sequencing and analysis of large populations.

Second generation sequencing systems

Second generation sequencing methods, also called next generation sequencing methods, were first introduced in 2005 and drastically changed the landscape of DNA sequencing. They increased throughput so much that genomes that previously took years to sequence could be sequenced within a matter of a few days to a few weeks. Here, I will

describe three next generation sequencing systems: 454 Genome Sequencer, Illumina Genome Analyzer and Ion Torrent's Personal Genome Machine. In general, these technologies have led to tremendous savings in cost and time required for genome analysis because of many reasons. First, the library preparation process has been greatly simplified. For these approaches, libraries are generally prepared by simply fragmenting, size selecting, amplifying and modifying genomic DNA to prepare it for sequencing. This has obviated the complicated, high-cost and time intensive cloning based approach to library preparation. Second, next generation sequencing instruments have lead to massively parallelized sequencing read generation, which has increased throughput considerably. Finally, availability of the human reference genome assembly (Lander et al. 2001; International Human Genome Sequencing Consortium 2004) has obviated the need for *de novo* assembly for computational analysis of sequencing read data, at least at the level of an entire genome. Consequently, sequencing reads are generally compared to the reference sequence to identify variants and provide a local genotype(s).

454 Genome Sequencer (Roche Applied Sciences): The first next generation sequencing instrument was commercialized by 454 Life Sciences, in 2005 (Margulies et al. 2005) and sequenced DNA using pyrosequencing (Nyrén et al. 1993). In this approach, randomly sheared adaptor-flanked DNA fragments are captured on micron-sized beads and amplified *in vitro* using PCR reactions done in emulsion droplets. After amplification, the emulsions are broken and the beads with amplified products are selectively recovered. Then, the amplification products are denatured and primed with a universal adaptor. These beads are then arrayed in individual wells on a fiber-optic slide. During each sequencing cycle, a single base is introduced into the wells. In the wells where extension happens, as defined by template sequences, a

pyrophosphate molecule is released, which through a series of steps is converted into a burst of light that is detected by a CCD imager coupled to the fiber optic array (Nyrén et al. 1993). This process is repeated many times to generate sequencing reads from single wells. Although high in throughput, 454 sequencing suffers from high error rates, particularly in homopolymeric regions where multiple incorporations happen in a single step leading to detection errors. Consequently, insertions and deletions in sequencing reads are the dominant errors with this technology.

Illumina Genome Analyzer: Commercialized by Illumina in 2006, the Genome Analyzer uses Solexa sequencing (Turcatti et al. 2008; Bentley et al. 2008). In this approach, a library of DNA fragments is deposited on a solid support and amplified using solid-phase amplification using bridge PCR (Adessi et al. 2000). Each fragment, localized to a unique physical location on the array, gives rise to a cluster of more than 1000 amplified copies, which generate a stronger signal for detection later. In the next step, DNA clusters are denatured and primed with a sequencing primer. Modified nucleotides are used during the extension process. These nucleotides, termed reversible terminators, are modified to contain a chemically cleavable fluorescent tag and a 3' blocking group. All four bases are modified with a different fluorescent dye. During each reaction cycle, all four nucleotides and a modified DNA polymerase are provided, leading to single incorporation and termination of extension at each cluster. The array is imaged, after which the 3' blocking group and the fluorescent dye are chemically cleaved, preparing the clusters for next cycle of extension. This cycle of extension, detection and cleavage is repeated many times to generate sequencing reads. Because of single incorporation in each step, this system does not have issues with homopolymeric DNA

sequences and is characterized by extremely high throughout. Illumina's HiSeq 2000 can generate 600 Gb of data per run. The latest system, HiSeq X combines 10 such sequencers and is projected to sequence more than 10,000 human genomes every year. These systems can also be used for paired-end sequencing, where sequencing reads are generated from the ends of a longer DNA fragment whose approximate size is known. It is, however, limited by short read lengths (commonly 100 bp) and high error rates because of dephasing.

Ion Torrent's Personal Genome Machine (PGM): The PGM uses a semiconductor chip called ion-sensitive field effect transistor (ISFET) that detects DNA incorporation based on pH changes in each well (Rothberg et al. 2011). During each cycle of extension, one dNTP is provided. In the wells where extension happens, a hydrogen ion is released as a byproduct and leads to local changes in pH, which is recorded by the semiconductor chip. By sequentially flowing bases, template molecules are sequenced. The PGM is relatively inexpensive and can be used for small scale sequencing projects.

Most of the second generation systems are very economical and high throughput in nature. Although affected by relatively higher error rates and short read lengths, high redundancy achieved through increased coverage can address these errors. However, short read lengths and the biases introduced due to amplification of template molecules preclude us from comprehensive genome analysis. A few third generation sequencing systems have been developed to produce much longer reads by sequencing single DNA molecules in real-time and addressing the issues with second generation systems. Here, I will describe two such systems: Single Molecule Read Time (SMRT) sequencing by Pacific Biosciences and nanopore based sequencing by Oxford Nanopore Technologies. I will also describe a single molecule sequencing

and physical mapping approach, called Optical Sequencing, which was developed by our lab in 2004.

Third generation sequencing systems

SMRT sequencing by Pacific Biosciences: Single molecule real time (SMRT) sequencing by Pacific Biosciences (Eid et al. 2009) uses zero-mode waveguide detectors (Levene et al. 2003) to study primer extension. Characterized by zeptolitre sequencing volumes, ZMWs allow the detection of single fluorescent dye tagged nucleotides upon incorporation in an immobilized DNA polymerase/template complex. During incorporation, the phosphate connected fluorescent dye is cleaved and washed out of the ZMW, allowing independent identification of next incorporated base. Also, all nucleotides have distinguishable fluorescent tags, which allow template sequencing. SMRT sequencing generates kilobase sized reads from single DNA molecules without halting the extension process.

Nanopore sequencing by Oxford Nanopore Technologies: The MiniION sequencing device by Oxford Nanopore Technologies uses electrical nanopore structures with proprietary enzymes to sequence single DNA molecules. Through each pore, current is passed by applying a voltage. As single stranded DNA is threaded though the pore, changes in current are observed specific to each base, and are used to identify DNA sequence (Branton et al. 2008). Kilobase-sized reads are currently reported using the MinION sequencer. However, it is affected by high error rates.

Optical sequencing: In 2004, our lab developed a single molecule DNA sequencing approach called optical sequencing (Ramanathan et al. 2004). Optical sequencing uses surface-immobilized DNA molecules to generate physical mapping information integrated with DNA sequence information at the physical markers. This is accomplished by presenting large single

DNA molecules analogously to Optical Mapping, nicking them with DNase, expanding the nicks to form gaps and finally, sequentially adding fluorochrome labeled nucleotides at nick sites *via* polymerase extension. In each cycle, only one type of fluorochrome tagged nucleotides are added as defined by template sequence, detected *via* fluorescence imaging, and photobleached to prepare the sites for next cycle of addition. Sequentially, incorporated fluorochromes are physically localized onto molecules that are also optically mapped allowing short strings of DNA sequence to be contextualized within a long double-stranded template. The combination of map and localized sequence reads builds the fundamental basis for *de novo* genome assembly, using very short sequence reads. This approach was built on comprehensive studies of consecutive or intermittent addition of fluorochrome tagged nucleotides during polymerase mediated primer extension, which demonstrated consecutive addition of up to 40 labeled nucleotides through specific combinations of polymerase type, fluor moiety and linker length (Ramanathan et al. 2005). This work laid the basis for development of Nanocoding (Jo et al. 2007). By combining physical mapping information with additional sequencing information, optical sequencing promises long range information from single DNA molecules that directly address short read length issues inherent to second generation sequencing methods.

In summary, the advances in second generation sequencing technologies have greatly accelerated our understanding of variation in normal and cancer genomes. Moving forward, it is anticipated that a combined strategy that uses low-error short reads from second generation sequencing systems and high-error long reads from third generation systems might lead us towards a comprehensive understanding of variation in normal and cancerous human genomes. Moreover, further improvements in read lengths and error rates from third

generation and other novel sequencing systems should overcome some of the current issues with genome analysis using DNA sequencing, which have been discussed in section 1.2.

3.1.2. Motivation

In this chapter, I will describe the use of paired-end DNA sequencing to study the multiple myeloma genomes described using Optical Mapping in chapter 2. There are three main objectives of this work: *i*) Identify somatic variation smaller than 3 kb in size, which is inaccessible to Optical Mapping; *ii*) Characterize breakpoints for optical structural aberrations, rearrangements and copy number changes identified from Optical Mapping at basepair resolution; and *iii*) serve as an orthogonal validation for our findings from Optical Mapping. Overall, this approach will lead to a comprehensive characterization of somatic variation in these multiple myeloma genomes. We are intently focused on determination of structural variation and rearrangements because although recent studies have helped us understand somatic variation in multiple myeloma at single nucleotide variant level (Chapman et al. 2011; Lohr et al. 2014), our understanding of structural variation in multiple myeloma and more generally, all cancers is very limited. A better understanding of prevalence of such variants might provide a better understanding of molecular mechanisms underlying tumor pathogenesis and chemotherapeutic response.

3.2. Study Design

We obtained whole genome paired-end sequencing data for Normal, drug sensitive MM-S and drug refractory MM-R samples, which were described previously in chapter 2. Using DNA sequencing data analysis tools and custom downstream analysis, we analyzed these tumor genomes for somatic single nucleotide, structural and copy number variants. Finally, we

integrated our findings from Optical Mapping with those from DNA sequencing based approaches to comprehensively identify and validate variation in these genomes. An overview of this approach is presented in fig. 3-1.

3.3. Materials and Methods

3.3.1. DNA preparation and data generation

For whole genome sequencing, genomic DNA was isolated from frozen cell samples using *DNA Isolation Kit for Cells and Tissues* (Roche) following the manufacturer's protocol. Paired-end Whole Genome Sequencing (WGS) was done by Beckman Coulter Genomics on Illumina HiSeq 2000 platform.

3.3.2. DNA sequencing data analysis

3.3.2.1. Trimming and adaptor removal: DNA sequencing data was cleaned (trimming and adaptor removal) using trim-galore (version 0.3.1) which is a wrapper script for cutadapt (cutadapt-1.2.1) (Martin 2011) and FastQC (version 0.10.1). Quality cutoff (q) of 20, error rate (e) of 0.1, stringency of 1 and minimum retained length (length) of 35 were used.

3.3.2.2. Alignment: The alignments of raw fastq reads to indexed human reference (NCBI Build 37) were done using Burrows-Wheeler Aligner (version bwa-0.7.5-a) (Li & Durbin 2009). The reads were aligned using default parameters with bwa aln and then paired using bwa sampe. The sam files thus generated were converted to bam format, merged for all lanes of one sample, coordinate sorted, duplicate marked, indexed and mate-fixed using picard tools (v1.84). The formatted bam files were realigned around indels and base recalibrated with GATK package (McKenna et al. 2010; DePristo et al. 2011) before being used for all downstream

analysis. For table 3-1, samtools (flagstat) (Li et al. 2009), picard (CollectInsertSizeMetrics) and GATK (DepthOfCoverage) were used to get the statistics.

3.3.2.3. SNP analysis: The Genome Analysis ToolKit (GATK) pipeline was used for SNP calling for Normal, MM-S and MM-R samples (McKenna et al. 2010; DePristo et al. 2011). To identify the commonly associated SNPs, the alignment (bam) files were manually visualized in IGV.

3.3.2.4. SNV/indel analysis using Strelka and comparison of SNVs with previous studies: Strelka (v1.0.10) (Saunders et al. 2012) was used with default settings to identify SNVs and small indels from the paired tumor-normal samples. To calculate SNV density for MM-S and MM-R samples, reference genome size of 2897.3 Mb (Build 37.1, Non-N bases in assembly) was used. The output SNVs and indels were annotated with SeattleSeq (Annotation 138). Exon-overlapping SNVs were identified and tabulated separately as synonymous and non-synonymous SNVs. Strelka calls somatic events at only those loci for which the Normal sample is homozygous reference. A visual inspection of bam files in IGV indicated that somatic events were called correctly. SNVs from MMRC cohort (Chapman et al. 2011) were lifted over to NCBI Build 37 coordinates and compared with SNVs from MM-S and MM-R samples using a custom perl script.

3.3.2.5. Structural variation analysis: Read pair (BreakDancer version 1.3.6) (Chen et al. 2009), split read (Pindel version 0.2.4o) (Ye et al. 2009) and read depth (CNVnator version 0.2.7) (Abyzov et al. 2011) based methods were used to identify structural variation in all three samples. breakdancer-max from BreakDancer package was used with default settings. For CNVnator, the genome was divided into 100 bp bins. It yielded an average read depth/bin of 61.649 +- 10.1042, 62.6684 +- 28.2886 and 100.998 +- 23.8661 for Normal, MM-S and MM-R

samples after GC correction. Pindel was run on all 3 samples concurrently using the default parameters except for –x 7, which defines the maximum event size as 524 kb. The output from these programs was analyzed and compared using custom scripts in R.

3.3.2.6. Comparison of deletions from Optical Mapping and DNA sequencing: For each of our three samples, the following steps were taken to compare Optical Mapping deletions to sequencing based deletion calls:

1) For Optical Mapping, deletions above 3 kb were considered after using the filters defined in Optical Mapping methods. During the process of manual curation, some deletions below 3 kb were added. However, majority of the deletions lie above 3 kb in size, as defined by automated structural variation calling from optical map assembly.
2) For BreakDancer, only the deletions which had more than 6 supporting reads and size between 3 kb and 250 kb were selected.
3) For CNVnator, deletions between 3 kb and 250 kb in size were selected. CNVnator calls regions as deletions if normalized read depth is <0.75.
4) For Pindel, only the deletions which had more than 6 unique supporting reads and size between 3 kb and 250 kb were selected.

The size filters for sequencing data were selected based on the size of deletions from Optical Mapping. The deletions from these methods were intersected using a custom perl script. The results were compiled and Venn diagram of overlap between different approaches was generated using 'VennDiagram' R package (Chen & Boutros 2011).

3.3.2.7. Comparison of insertions from Optical Mapping and DNA sequencing: For each of our three samples, the following steps were taken to compare Optical Mapping insertions to sequencing based insertion calls:

1) For Optical Mapping, insertions above 3 kb were considered after using the filters defined in Optical Mapping methods. During the process of manual curation, some insertions below 3 kb were added. However, majority of the insertions lie above 3 kb in size, as defined by automated structural variation calling from optical map assembly.
2) For CNVnator, duplications between 3 kb and 250 kb in size were selected. CNVnator calls regions as duplications if normalized read depth is >1.25.
3) Pindel generates a list of Long Insertions (LI) for which it does not identify either the size or the sequence of inserted DNA. It generates the approximate reference basepair location of the insertion. We used only those LI calls where the number of supporting reads >=6.

The size filters for sequencing data were selected based on the size of insertions from Optical Mapping. The insertion/duplication calls thus obtained were intersected using a custom perl scripts and Venn diagrams of overlap between different approaches was generated using 'VennDiagram' R package.

3.3.2.8. Identification of somatic deletions shared between MM-S and MM-R samples: For deletions less than 3 kb in size, only DNA sequencing based deletion calls were considered. For deletions greater than 3 kb in size, deletion calls from Optical Mapping and DNA sequencing were integrated to generate the final list of somatic deletions shared between MM-S and MM-R samples.

1) Optical Mapping: A Custom perl script was used to identify deletions found in MM-S and MM-R but not in Normal sample.
2) DNA Sequencing: For CNVnator and BreakDancer, the minimum deletion call is ~300 bp in size, while for Pindel, deletions range from 1 bp - 524 kb. For this reason, deletion calls from DNA sequencing data were split into two categories by size: 10 - 400 bp and 400 bp - 250 kb.
    - 10 - 400 bp: Pindel output was filtered for 10 - 400 bp long deletion calls. Number of unique supporting reads in Normal, MM-S and MM-R was =0, >=10 and >=10, respectively.
    - 400 bp – 250 kb: Pindel output was filtered such that number of unique supporting reads in Normal, MM-S and MM-R samples was <=2, >=10 and >=10, respectively. The output was compared to and annotated with BreakDancer and CNVnator output.

Finally, results from Optical Mapping and DNA sequencing were combined. The final list of somatic deletions shared between MM-S and MM-R samples was annotated with overlapping genes and exons using intersectBed utility from bedtools-2.18.0 package (Quinlan & Hall 2010).

3.3.2.9. Identification of somatic deletions unique to MM-R sample: For deletions less than 3 kb in size, only DNA sequencing based called were considered. For deletions greater than 3 kb in size, deletion calls from Optical Mapping and DNA sequencing were integrated to generate the final list of somatic deletions shared between MM-S and MM-R samples.

1) Optical Mapping: A Custom perl script was used to identify deletions found in MM-R sample but not in Normal and MM-S samples.
2) DNA sequencing:
    - 10-400 bp: Only Pindel output was considered for these calls. A deletion call was considered unique to the MM-R sample if number of unique supporting reads in Normal, MM-S and MM-R samples was =0, =0, and >=10, respectively. These filters were chosen empirically after using different filter sets and visualizing the output in Integrative Genomics Viewer (IGV) (Robinson et al. 2011).
    - 400 bp onwards: Output from BreakDancer, CNVnator and Pindel was considered for deletions >400 bp in size. In a first pass, a deletion call from Pindel was considered unique to MM-R sample if number of unique supporting reads in Normal, MM-S and MM-R samples was <=2, <=2 and >=10, respectively. The filtered output was compared to and annotated with BreakDancer and CNVnator output. Some of the original Pindel calls were eliminated because they weren't validated by other methods.

Finally, results from Optical Mapping and DNA sequencing were combined and the final list of somatic deletions in MM-R sample was annotated with overlapping genes and exons.

3.3.2.10. Somatic Copy Number Analysis

Control-FREEC (Boeva et al. 2012) was used for somatic copy number alterations (SCNA) analysis using a window size of 50 kb and a step size of 10 kb. The results were plotted using Circos (circos-0.64) (Krzywinski et al. 2009) and annotated using UCSC genome browser (http://genome.ucsc.edu/index.html) (Kent et al. 2002).

3.4. Results

3.4.1. Summary statistics of data collection

A summary of paired end sequencing data has been presented in table 3-1. We collected sequencing data representing an averaged genome-wide coverage of 56.98, 68.41 and 94.78 fold for Normal, MM-S and MM-R samples, respectively. The data were of good quality, as indicated by high percentage of properly paired reads (>90%) and low percentage of duplicate reads (4.66 % - 12.33 %).

3.4.2. Single Nucleotide Polymorphism (SNP) analysis

Previous genome-wide association studies (GWAS) have identified 8 inherited susceptibility markers that are associated with multiple myeloma (Broderick et al. 2012; Weinhold et al. 2013; Chubb et al. 2013). The objective of our SNP analysis was to identify which of these 8 loci are present in our patient samples. Upon analysis, we identified 7 out of 8 of these markers in our samples in a homozygous or a heterozygous state. A summary is provided in table 3-2. Among these is a SNP in gene CCND1 at 11q13.3 (c.870G>A, rs9344), which has previously been implicated as a risk factor for t(11;14)(q13;q32) multiple myeloma (Weinhold et al. 2013).

3.4.3. Single Nucleotide Variation (SNV) analysis

Like in other cancers, the genome in multiple myeloma is known to acquire somatic point mutations (also called SNVs) with disease progression. Recent large scale sequencing studies, where the authors have identified SNVs by comparing the tumor samples at presentation or after treatment to corresponding normal samples, have highlighted the patterns of point mutations in multiple myeloma (Chapman et al. 2011; Lohr et al. 2014).

The objective of our SNV analysis was to compare the tumor samples (MM-S and MM-R) to the Normal sample and identify *i*) SNVs acquired in MM-S sample and retained in MM-R sample; and *ii*) SNVs acquired in and unique to the MM-R sample. We hypothesized that some SNVs acquired at MM-S stage might contribute to multiple myeloma pathogenesis and some that are novel to the MM-R stage might explain the onset of refractory disease in this patient. For this analysis, we used Strelka - a somatic variant caller that can be used to detect SNVs and small indels from aligned sequencing reads of matched tumor-normal samples (Saunders et al. 2012).

3.4.3.1. SNVs reveal increased mutational complexity with tumor progression: Based on our analysis, we identified 10,224 and 13,511 SNVs in MM-S and MM-R samples, respectively, when compared to the Normal sample. This yields tumor-specific point mutations rates of 3.53 and 4.52 per million bases for MM-S and MM-R samples, respectively, indicating an increase in mutational burden with multiple myeloma progression (Table 3-3). Although a bit higher than the average tumor-specific point mutation rate from the MMRC cohort (2.9 per million bases) (Chapman et al. 2011), it can be explained by higher sequencing depth or increased incidence of somatic mutations in our samples.

3.4.3.2. Functional annotation of SNVs: Next, we annotated the identified SNVs using SeattleSeq, a web-based tool that provides functional annotations of SNVs with dbSNP IDs, gene names and accession numbers, variation functions (e.g. missense), protein positions and amino-acid changes, PolyPhen predictions, and clinical association using data pooled from many sources. Fig. 3-2 shows a distribution of all SNVs identified in MM-S and MM-R samples across different genomic elements. The annotation revealed that the MM-S sample had 27

synonymous and 90 nonsynonymous SNVs (Table 3-4) and the MM-R sample had 33 synonymous and 101 nonsynonymous SNVs (Table 3-5). We further compared the nonsynonymous SNVs from MM-S and MM-R samples and found that only 60 were shared between both samples. Novel SNVs in the MM-R sample (41 in number) corroborate our overall finding of the increase in mutational burden with tumor progression. Notable are truncating mutations in MYB and BRCA2, which are unique to the MM-S sample. Consistent with previous findings from primary tumor samples (Politou et al. 2006; Oerlemans et al. 2008; Lü et al. 2009; Ri et al. 2010), we did not observe any PSMB5 proteasomal mutations in MM-S and MM-R samples.

3.4.3.3. Comparison of SNVs with previous multiple myeloma sequencing studies reveals no overlap: We next wanted to check if any of the SNVs from our analysis overlapped with SNVs identified in the MMRC cohort (Chapman et al. 2011) or another recent sequencing study of a multiple myeloma patient (Egan et al. 2012). Not surprisingly, we did not find any shared SNVs between our work and these previous studies. Upon further inspection, we did not find any SNVs that were shared between the MMRC cohort and the other study either (Egan et al. 2012). This further supports the notion of widespread inter-tumor heterogeneity in multiple myeloma and that disparate mutations possibly lead to deregulation of common downstream pathways, thereby contributing to multiple myeloma pathogeneis (Chapman et al. 2011; Lohr et al. 2014).

3.4.3.4. Copy number neutral loss of heterozygosity (LOH) regions: Based on our work with Optical Mapping, we had some indication that these tumor genomes harbor a few copy number neutral segments with LOH. Such regions have been commonly reported for hematological

malignancies (O'Keefe et al. 2010). We reasoned that irrespective of copy number, LOH regions should show a higher density of SNVs because in these regions, many loci which were heterozygous in the Normal sample, will present as homozygous in tumor samples and consequently, will be identified as SNVs in our analysis using Strelka. By then observing the depth of coverage across these regions, we could distinguish copy number neutral LOH (depth similar to overall depth) from copy number loss related LOH (depth much lower than overall depth). Accordingly, we identified four regions spanning a total of 215 Mb on q arms of chr1, chr5 and chr14 in MM-R sample that presented with copy number neutral LOH (Fig. 3-3C, Table 3-6). Please note that Fig. 3-3 aggregates many results from this work, and I will refer to this figure many times in the text.

3.4.4. Structural variation analysis

Although structural variation analysis using DNA sequencing data is marred by a very high false positive rate and a yet unknown false negative rate (Abel & Duncavage 2013), we reasoned that an integrative approach might provide us useful insights about structural variation in these samples. Hence, we selected a number of DNA sequencing based structural variation calling approaches, as discussed in following sections, and integrated common findings from these approaches with those from Optical Mapping structural variation analysis to learn more about structural variation in multiple myeloma.

For paired end sequencing data, a number of computational approaches and algorithms have been developed to identify structural variants. In general, they are based on the analysis of three features of DNA sequencing data: read pairs, read depth and split reads. BreakDancer (read-pair based; Chen et al. 2009) uses paired reads with discordant insert sizes or orientation

to identify structural variants that include deletions, inversions and intra- and inter-chromosomal translocations. CNVnator (read-depth based; Abyzov et al. 2011) uses read-depth across the genome for copy number variation (CNV) discovery and genotyping. CNVnator partitions the genome into non-overlapping bins of equal size, calculates the mean depth of coverage for each bin, and after correcting for GC and other biases, divides the genome into regions of constant copy number. It can identify amplifications and deletions. Finally, Pindel (split-read based; Ye et al. 2009) uses a pattern growth approach to identify the breakpoints of structural variants, at which the sequenced genome differs from the reference genome, based on the paired-end short reads. It can identify small insertions, breakpoints of large insertions, deletions, inversions and tandem duplications.

3.4.4.1. Analysis of deletions

While Optical Mapping effectively identifies structural variation, it does not characterize the structural variants at basepair resolution. Thus, we used DNA sequencing based structural variation identification and assembly methods to place sequence data atop structural variants. Our strategy serves two purposes for comprehensiveness: *i*) validation of Optical Mapping deletions; *ii*) basepair level resolution of deletions identified from Optical Mapping and identification of deletions that lie below Optical Mapping resolution.

3.4.4.1.1. Validation of Optical Mapping deletion calls: We compared the deletion calls from Optical Mapping to those from DNA sequencing data analyzed using read-pair (BreakDancer), split-read (Pindel) and read-depth (CNVnator) based approaches. Close to 80% of the Optical Mapping deletions were captured by one of more sequencing based methods. More

specifically, we validated 108/139 (77.7%), 114/149 (76.5%) and 138/176 (78.4%) deletion calls from Optical Mapping in Normal, MM-S and MM-R samples, respectively (Fig. 3-4).

We further analyzed the deletions that were identified only by Optical Mapping in MM-R sample (38/176) and found that majority of them overlapped with segmental duplications (13/38) or repeat DNA sequences (11/38). A few (5/38) were less than 3 kb in length or had non-matching breakpoints in the same region (2/38) and hence, were not captured in this overlap. Overall, this points to high accuracy of deletion calling and the ability to identify variation in duplication/repeat rich regions using Optical Mapping.

However, we also observed that Optical Mapping missed some deletions that were captured by all DNA sequencing based approaches, which can be explained by three factors. First, the Optical Mapping variation calling algorithm missed some hemizygous deletions because in the current implementation, it compares just one representation of genome assembly to *in silico* reference maps. Second, the filters that are used to reduce false positives during variation calling from optical map data also eliminate some true calls. Lastly, we noticed that some deletions from sequencing data involved more complex rearrangements and hence, were catalogued in a different category (COMPLEX) by Optical Mapping variation calling.

3.4.4.1.2. All somatic deletions in MM-S and MM-R samples: Using a combination of Optical Mapping and DNA sequencing analysis, we identified many somatic deletions that we have divided into two categories: shared between MM-S and MM-R samples, and novel to just the MM-R sample. The genes affected by these deletions have been summarized in Table 3-7.

<u>Somatic deletions shared between MM-S and MM-R samples:</u> We identified 38 somatic deletions larger than 400 bp, which range in size from 497 bp to 192 kb (Table 3-8). Of these, 10

deletions overlap exons. Furthermore, we identified 6 small deletions (10 – 400 bp) that overlap exons (Table 3-8). Among these is a 73 bp deletion associated with a TP53 exon. As discussed earlier, MM-S and MM-R samples also show del(17p), which indicates that both copies of TP53 are inactivated in these samples.

Somatic deletions novel to MM-R: We identified another 27 deletions larger than 400 bp, which range from 445 bp to 290 kb in size and are unique to the MM-R sample (Table 3-9). Of these, 7 deletions overlap exons. Furthermore, we identified 2 small deletions (10 – 400 bp) that overlap exons (Table 3-9).

3.4.4.2. Analysis of insertions

While deletions are relatively straightforward to identify using sequencing based approaches, identifying insertions is rather difficult because most of the sequencing based tools, except *de novo* assembly, use an entirely reference based approach to detect variation. We compared the insertions from Optical Mapping to long insertions from Pindel and duplications from CNVnator. Pindel generates insertion breakpoints with no information about the length or structure of the insertions, whereas CNVnator generates a list of duplications with no structural or contextual information. As expected, a smaller percentage (24 – 40 %) of Optical Mapping insertions had an overlapping Pindel or CNVnator call (Fig. 3-5). It is also interesting to note that thousands of duplications/insertions, which have no intersecting calls from other approaches, have been called by sequencing based methods. Most likely, a large number of these calls are false positives.

3.4.5. Somatic copy number alteration analysis

DNA Sequencing data can be used to identify somatic copy number changes from paired tumor-normal samples. We used a software package called Control-FREEC because it detects somatic copy-number changes from paired samples by automatically computing, normalizing, segmenting copy number and beta allele frequency (BAF) profiles, then calling copy number alterations (Boeva et al. 2012). Using Control-FREEC, we compared the MM-S and MM-R tumor samples to the Normal sample and identified large-scale copy number gains and losses, as were previously documented with Optical Mapping HMM coverage analysis in chapter 2. Figs. 3-6 and 3-7 present somatic copy number profiles for MM-S and MM-R samples, respectively. Visual inspection of these copy number profiles indicated concordance with HMM copy number analysis from Optical Mapping alignment data. A more rigorous comparison of DNA sequencing based somatic copy number alteration analysis with Optical Mapping based HMM coverage analysis revealed that for genomic segments larger than 500 kb, 97% of the genome was assigned concordant copy number states by both approaches (Fig. 3-3D and 3-3E).

As described in methods, we used a 10 kb window for copy number analysis using DNA sequencing data. Consequently, we got approximate copy number breakpoints, which were within 10 kb of exact breakpoints. We reasoned that if we use these approximate breakpoint locations and look up structural variation output calls near these locations from BreakDancer and Pindel, both of which generate basepair level resolved structural variants, we might be able to resolve some structural rearrangements that were not previously identified using Optical Mapping (summarized in table 2-7). We also reasoned that if both breakpoints of these structural variation calls coincided with copy number breakpoints identified from our analysis, it would be a strong indicator of the rearrangement being true and not a data analysis artifact.

Upon doing this manual analysis, we were able to resolve genomic rearrangements underlying 6 more copy number aberrant regions. These rearrangements included translocations between chr1/chr6, chr5/chr7, chr8/chr9, and chr14/chr21 (Fig. 3-3F). Here, I describe two translocations in more detail.

1q and 6p amplification from t(1;6): Amplifications of 1q region are commonly reported for multiple myeloma, with 1q21.1-1q23.3 identified as the minimally amplified region (Walker et al. 2010). We observed the minimal amplification in our sample (Fig. 2-7, 3-3, 3-7), with one of the amplification breakpoints associated with chr6 in a translocation, also resulting in a 51 Mb amplification on p-arm of chr6. We also noticed that 1q amplification occurred late during tumor progression because it was not observed in MM-S sample.

8p and 9p loss from t(8;9): A deletion at 8p locus is reported for 29% patients, with 8p23.1-8p12 characterized as the minimally deleted region (Walker et al. 2010). We observed a larger deletion of ~36 Mb from 8p32.1 to 8p12 in our samples. Furthermore, the deletion breakpoint is connected to another deletion breakpoint on chr9 through a translocation that leads to additional loss of ~32 Mb at 9p locus (Figs. 2-7, 3-3, 3-7).

Additionally, by doing a similar look-up for structural rearrangements previously described using Optical Mapping, we were able to resolve their breakpoints to basepair resolution. Overall, we were able to describe structural rearrangements that explained 31 out of 37 copy number breakpoints. A summary of these rearrangements is provided in Table 3-10 and illustrated in Fig. 3-3F.

3.4.6. Putting it all together - Optical Mapping and DNA sequencing complement each other, leading to comprehensive understanding of genome structure

By combining analysis from Optical Mapping and DNA sequencing data, we were able to piece together the structure of many chromosomes. Here, we describe the structure of two chromosomes in more detail.

3.4.6.1. Detailed analysis of chr5 in MM-R sample: Chr5 presented with two regions of copy number loss: 70.74 - 78.41 Mb and 96.63 - 134.01 Mb (Fig. 3-8). From Optical Map assembly, we observed a chromosomal end at 70.74 Mb, indicating that the entire region following this breakpoint was lost *via* truncation in one copy. Also, single nucleotide polymorphism (SNP) analysis indicated loss of heterozygosity (LOH) following 70.74 Mb for both, normal and reduced copy number regions, indicating that regions of normal copy number resulted from copy number neutral LOH. Optical Map assembly presented another breakpoint at 96.63 Mb, detailed analysis of which revealed a 18 Mb long tandem duplication of 78.41 – 96.63 Mb region, explaining copy number neutral LOH in this region. Based on sequencing data analysis, the second copy of 134.01 Mb – End region was found to be associated in a translocation with chr7. Finally, Optical Map assembly revealed a t(5;11)(q35.3;q23.3) unbalanced translocation, which is associated with loss of ~900 kb (180.03 – 180.91 Mb) at q-ter of chr5 and amplification of ~15.5 Mb from q-ter of chr11. It is interesting to note that the 70.74 Mb breakpoint region overlaps a segmental duplication and the 96.63 Mb breakpoint overlaps a LINE repeat, thereby obscuring these regions to full analysis using DNA sequencing data alone. However, we were able to resolve and understand highly rearranged structure of chr5 using information from optical map assembly, copy number analysis and DNA sequencing data.

3.4.6.2. Detailed analysis of chr7 in MM-S and MM-R samples: In both MM-S and MM-R samples, chr7 presented with a 26 Mb region of copy number 4 (20.87 Mb – 46.92 Mb) and a

53.5 Mb region of copy number 1 (67.26 Mb – 120.75 Mb) (Fig. 3-3, 3-9). We identified two separate translocations with chr16 at both breakpoints of the copy number 4 region, which explain copy number gains observed on chr7 and chr16. Furthermore, the 67.26 Mb breakpoint revealed an inverted duplication that leads to chromosomal truncation. Finally, the region following 120.75 Mb breakpoint on chr7, which shows normal copy number of 2, is associated in a translocation with chr12, also leading to loss of ~21 Mb at p-ter of chr12, which is a common multiple myeloma loss (Walker et al. 2010).

3.4.7. Impact on multiple myeloma biology

The analysis presented here, although limited to a single individual, offers unique biological insights that may translate into novel strategies for targeted interventions. Individual genes mutated in both MM-S and MM-R samples may represent clonal drivers associated with core mechanisms of multiple myeloma oncogenesis and/or acquisition of drug resistance. These clonal events constitute legitimate therapeutic targets, particularly in the light of recent data demonstrating the limits, and potential dangers, associated with targeting subclonal mutations (e.g. mutant *BRAF*) (Lohr et al. 2014). Several events have been established previously as important progression factors in myeloma pathogenesis (e.g., *TP53* and *CDKN2C* loss) (Leone et al. 2008; Kuehl & Bergsagel 2012). *PIK3R1* aberrations underscore the importance of the PI3K pathway activation in multiple myeloma (Harvey & Lonial 2007). IGF2BP2 mutations affect insulin growth factor-2 (IGF-2) translation and thus growth control through an IGFR-controlled pathway that is currently the focus of intense pre-clinical development in myeloma (Menu et al. 2009). Other genes are less well studied in multiple myeloma and may provide novel potential targets or regulatory pathways. ELL is an essential

cofactor of the super-elongation complex (SEC), a central node of transcriptional elongation checkpoint control and key mutational target in myeloid and mixed-lineage leukemias (Smith et al. 2011). TCL1 is a regulator of apoptosis implicated in B-chronic lymphocytic leukemia as well as T-cell lymphomas (Pekarsky et al. 2004). ASXL3 belongs to a family of epigenetic regulators whose genetic loss has been implicated in the myelodysplastic syndromes and myeloid leukemias (Shih et al. 2012). CAMK2D is a subunit of calmodulin-dependent kinase II, an essential regulator of Ca^{2+}-dependent signal transduction. Interestingly, prior studies have implicated calmodulin-dependent pathways in proteasome inhibitor-resistant, constitutive NFkB activity in multiple myeloma (Berchtold et al. 2005; Markovina et al. 2010). Our analysis has therefore pinpointed novel potential targets that merit validation in larger cohorts of multiple myeloma patients prior to functional experimentation using appropriate model systems.

Intriguingly, genes uniquely affected in the MM-R sample include cell cycle regulators (CCNG2), mitotic checkpoint genes (ZWILCH) as well as a transcription factor (MYBL1) expressed specifically in centroblasts (Golay et al. 1998), a putative pre-plasmablastic cell-of-origin for multiple myeloma within the germinal center reaction (Kuehl & Bergsagel 2012). These findings raise the possibility that disease progression and/or acquisition of drug resistance in multiple myeloma may be associated with plasmacytic maturation arrest or de-differentiation to earlier stages of B cell ontogenesis. Thus effective management of end-stage myeloma may necessitate approaches that promote mitotic quiescence or cell cycle exit and terminal plasmacytic differentiation.

3.5. Conclusions

In conclusion, we have comprehensively identified and characterized single nucleotide, structural and copy number variation in a multiple myeloma genome. We have also characterized most of the structural rearrangements that lead to copy number aberrations and consequently, decoded parts of highly complex genome structure in these cancer genomes. Using our serial analysis, we have identified many somatic variants which are uniquely observed in the drug refractory tumor sample (MM-R). These additional variants, which include SNVs, deletions and large scale copy number changes resulting from genomic rearrangements, provide evidence for increasing mutational complexity with tumor progression at all size scales. Also, since they are observed only in the drug refractory sample, some of them might contribute to the mechanisms underlying drug resistance and can be evaluated in future studies for their functional impact.

By further characterization of Optical Structural Aberrations, we have shown how Optical Mapping can identify structural variation associated with repeat rich regions of the genome, and consequently, delineate genomic structure and rearrangements in these tough to analyze regions. High rate of validation of deletions (~80%) and copy number changes (~97%) indicates high accuracy of Optical Mapping based variation and copy number identification.

Finally, we have demonstrated how an integrated approach using Optical Mapping and DNA sequencing leverages the unique advantages of these genome analysis systems, and provides us a platform to rigorously analyze highly rearranged cancer genomes.

3.6. Bibliography

Abel, H.J. & Duncavage, E.J., 2013. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer genetics*, 206(12), pp.432–40.

Abyzov, A. et al., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6), pp.974–84.

Adessi, C. et al., 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research*, 28(20), p.E87.

Ansorge, W. et al., 1986. A non-radioactive automated method for DNA sequence determination. *Journal of biochemical and biophysical methods*, 13(6), pp.315–23.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–9.

Berchtold, C.M. et al., 2005. Perillyl alcohol inhibits a calcium-dependent constitutive nuclear factor-kappaB pathway. *Cancer research*, 65(18), pp.8558–66.

Boeva, V. et al., 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(3), pp.423–5.

Branton, D. et al., 2008. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10), pp.1146–53.

Broderick, P. et al., 2012. Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nature genetics*, 44(1), pp.58–61.

Chapman, M.A. et al., 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), pp.467–72.

Chen, H. & Boutros, P.C., 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC bioinformatics*, 12(1), p.35.

Chen, K. et al., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9), pp.677–81.

Chubb, D. et al., 2013. Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nature genetics*, 45(10), pp.1221–5.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), pp.491–8.

Dovichi, N.J., 1997. DNA sequencing by capillary electrophoresis. *Electrophoresis*, 18(12-13), pp.2393–9.

Egan, J.B. et al., 2012. Whole genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution and clonal tides. *Blood*. 120(5), pp.1060-6.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910), pp.133–8.

Golay, J. et al., 1998. The A-Myb transcription factor is a marker of centroblasts in vivo. *Journal of immunology (Baltimore, Md. : 1950)*, 160(6), pp.2786–93.

Harvey, R.D. & Lonial, S., 2007. PI3 kinase/AKT pathway as a therapeutic target in multiple myeloma. *Future oncology (London, England)*, 3(6), pp.639–47.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–45.

Jo, K. et al., 2007. 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), pp.2673–8.

Kent, W.J. et al., 2002. The Human Genome Browser at UCSC. *Genome Research*, 12(6), pp.996–1006.

Krzywinski, M. et al., 2009. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), pp.1639–45.

Kuehl, W.M. & Bergsagel, P.L., 2012. Molecular pathogenesis of multiple myeloma and its premalignant precursor. *The Journal of clinical investigation*, 122(10), pp.3456–63.

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

Leone, P.E. et al., 2008. Deletions of CDKN2C in multiple myeloma: biological and clinical implications. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(19), pp.6033–41.

Levene, M.J. et al., 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, N.Y.)*, 299(5607), pp.682–6.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60.

Lohr, J.G. et al., 2014. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell*, 25(1), pp.91–101.

Lü, S. et al., 2009. Different mutants of PSMB5 confer varying bortezomib resistance in T lymphoblastic lymphoma/leukemia cells derived from the Jurkat cell line. *Experimental hematology*, 37(7), pp.831–7.

Luckey, J.A. et al., 1990. High speed DNA sequencing by capillary electrophoresis. *Nucleic acids research*, 18(15), pp.4417–21.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80.

Markovina, S. et al., 2010. Bone marrow stromal cells from multiple myeloma patients uniquely induce bortezomib resistant NF-kappaB activity in myeloma cells. *Molecular cancer*, 9, p.176.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), p.10.

Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp.560–4.

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), pp.1297–303.

Menu, E. et al., 2009. The role of the insulin-like growth factor 1 receptor axis in multiple myeloma. *Archives of physiology and biochemistry*, 115(2), pp.49–57.

Morozova, O. & Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), pp.255–64.

Nyrén, P., Pettersson, B. & Uhlén, M., 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical biochemistry*, 208(1), pp.171–5.

O'Keefe, C., McDevitt, M.A. & Maciejewski, J.P., 2010. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood*, 115(14), pp.2731–9.

Oerlemans, R. et al., 2008. Molecular basis of bortezomib resistance: proteasome subunit beta5 (PSMB5) gene mutation and overexpression of PSMB5 protein. *Blood*, 112(6), pp.2489–99.

Pekarsky, Y. et al., 2004. Tcl1 as a model for lymphomagenesis. *Hematology/oncology clinics of North America*, 18(4), pp.863–79, ix.

Politou, M. et al., 2006. No evidence of mutations of the PSMB5 (beta-5 subunit of proteasome) in a case of myeloma with clinical resistance to Bortezomib. *Leukemia research*, 30(2), pp.240–1.

Prober, J.M. et al., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (New York, N.Y.)*, 238(4825), pp.336–41.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–2.

R: a language and environment for statistical computing | GBIF.ORG. Available at: http://www.gbif.org/resources/2585 [Accessed August 19, 2014].

Ramanathan, A. et al., 2004. An integrative approach for the optical sequencing of single DNA molecules. *Analytical biochemistry*, 330(2), pp.227–41.

Ramanathan, A., Pape, L. & Schwartz, D.C., 2005. High-density polymerase-mediated incorporation of fluorochrome-labeled nucleotides. *Analytical biochemistry*, 337(1), pp.1–11.

Ri, M. et al., 2010. Bortezomib-resistant myeloma cell lines: a role for mutated PSMB5 in preventing the accumulation of unfolded proteins and fatal ER stress. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, 24(8), pp.1506–12.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature biotechnology*, 29(1), pp.24–6.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–52.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), pp.441–8.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–7.

Saunders, C.T. et al., 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, 28(14), pp.1811–7.

Shih, A.H. et al., 2012. The role of mutations in epigenetic regulators in myeloid malignancies. *Nature reviews. Cancer*, 12(9), pp.599–612.

Smith, E., Lin, C. & Shilatifard, A., 2011. The super elongation complex (SEC) and MLL in development and disease. *Genes & development*, 25(7), pp.661–72.

Smith, L.M. et al., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071), pp.674–9.

Swerdlow, H. et al., 1991. Three DNA sequencing methods using capillary gel electrophoresis and laser-induced fluorescence. *Analytical chemistry*, 63(24), pp.2835–41.

Swerdlow, H. & Gesteland, R., 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic acids research*, 18(6), pp.1415–9.

Turcatti, G. et al., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4), p.e25.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.

Walker, B. a et al., 2010. A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, 116(15), pp.e56–65.

Weinhold, N. et al., 2013. The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nature genetics*, 45(5), pp.522–5.

Ye, K. et al., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21), pp.2865–71.

3.7. Figures

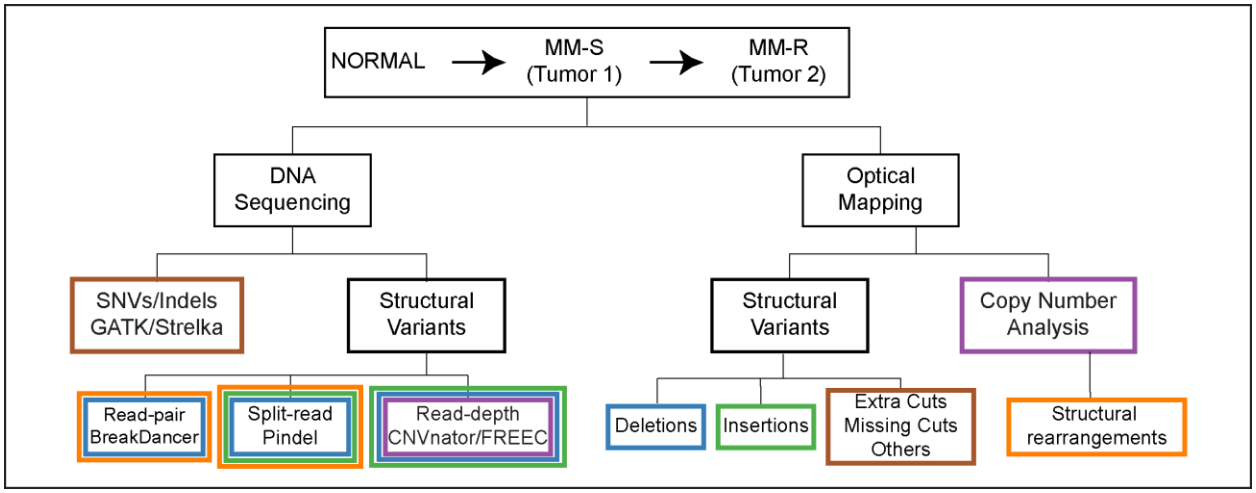

Fig. 3-1: An overview of our cancer genome analysis pipeline using DNA sequencing and Optical Mapping. Colored outlines highlight different variation types analyzed by integrating data from both approaches. E.g. deletions from Optical Mapping (blue outline) were analyzed along with deletions from BreakDancer, Pindel and CNVnator.

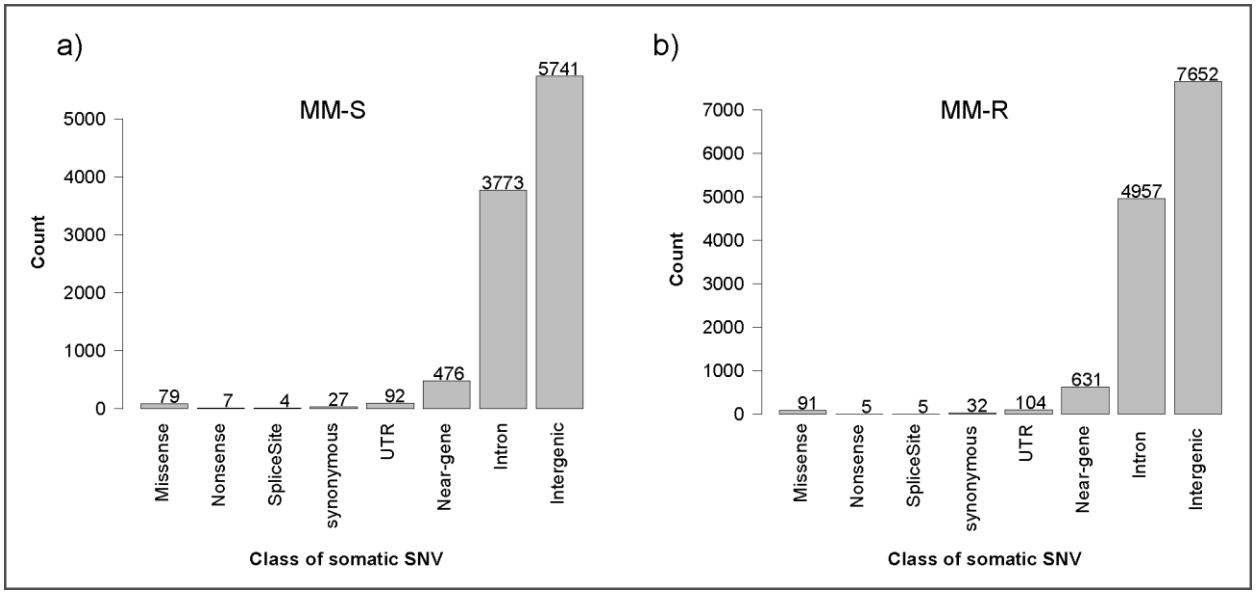

Fig. 3-2: Distribution of SNVs in MM-S (a) and MM-R (b) samples across various genomic elements.

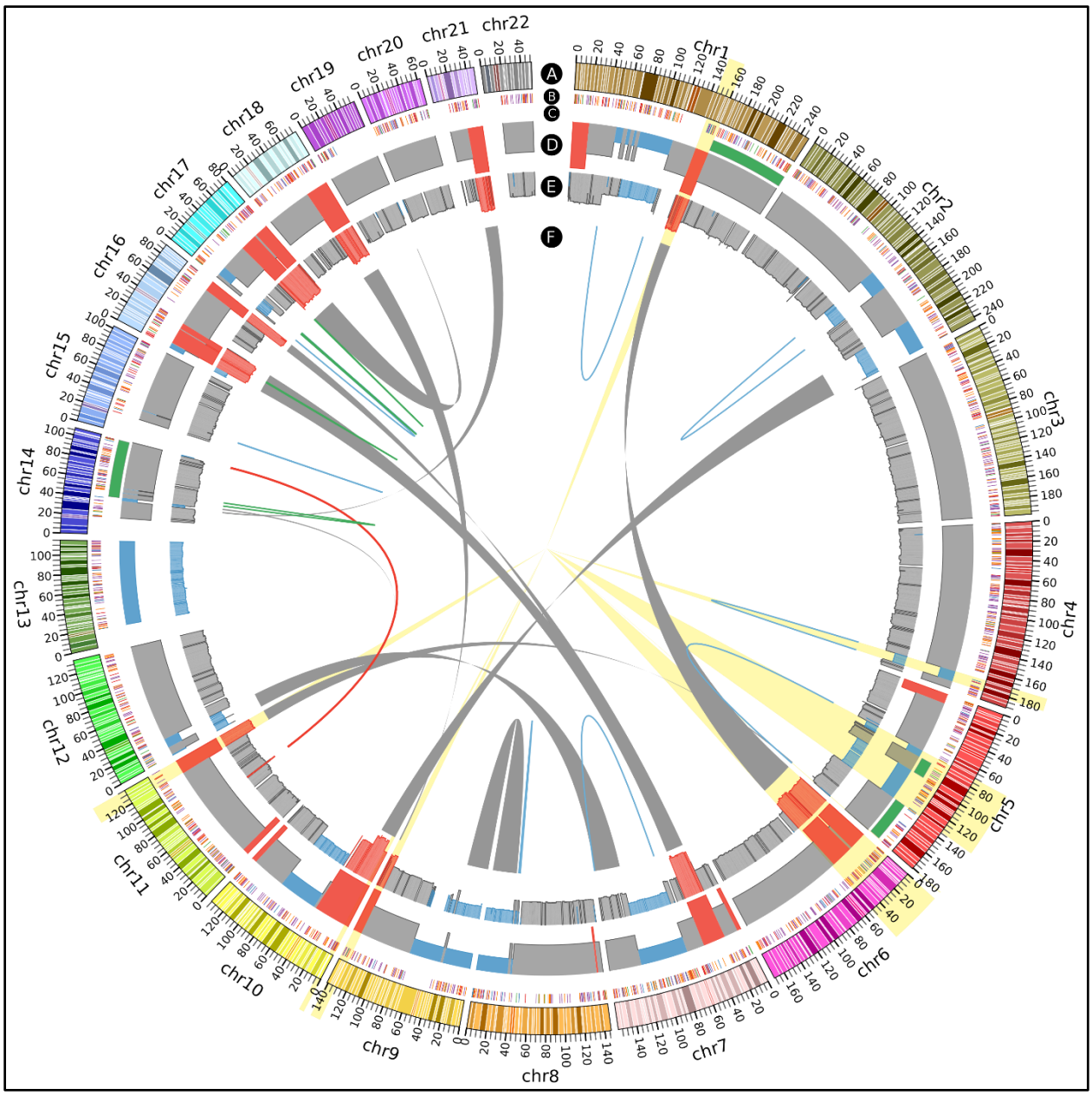

Fig. 3-3: Circos plot of structural variants, copy number (CN) variants and genomic rearrangements in multiple myeloma. Tracks are as follows: A) Chromosomal ideograms showing chr1 to chr22 in clockwise orientation (chr8 reversed; chrX & chrY excluded for clarity). B) Structural variants in MM-R sample identified using Optical Mapping (insertions, deletions, extra cuts, missing cuts and other complex rearrangements). C) CN neutral regions with loss of heterozygosity (LOH), identified from SNP analysis of DNA sequencing data. D) Somatic CN aberrations from optical map coverage analysis of MM-R *vs.* Normal sample: High CN (red), normal CN (grey) and low CN (blue) states are shown. E) Somatic CN aberrations from DNA sequencing data for MM-R *vs.* Normal sample using FREEC package: High CN (red), normal CN (grey) and low CN (blue) regions are shown. F) Links represent genomic rearrangements that coincide with and lead to CN changes identified from Optical Map assembly and/or DNA sequencing data. Grey links connect translocation breakpoints; red link represents canonical

t(11;14) translocation in multiple myeloma; blue links show large interstitial and terminal deletions; green links represent inversions. Light yellow highlights (background; center to outside the circle) describe CN aberrations and associated structural variants not observed in MM-S sample, and unique to MM-R sample.

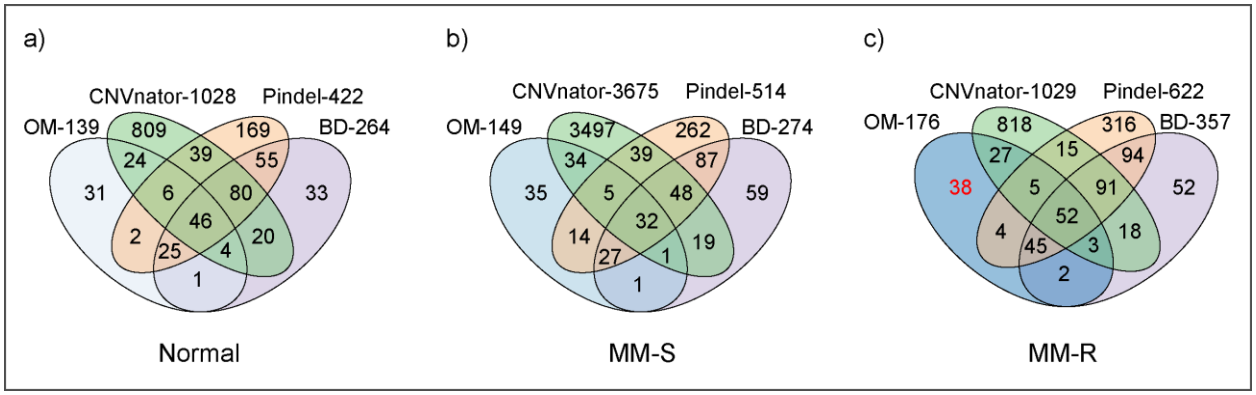

Fig. 3-4: Intersection Venn diagrams of deletion calls from Optical Mapping (OM) with deletion calls from DNA sequencing based structural variation analysis methods BreakDancer (BD), CNVnator and Pindel for Normal (a), MM-S (b) and MM-R (c) samples. The text at top labels the approach followed by total number of deletions identified by that approach using filters defined in text.

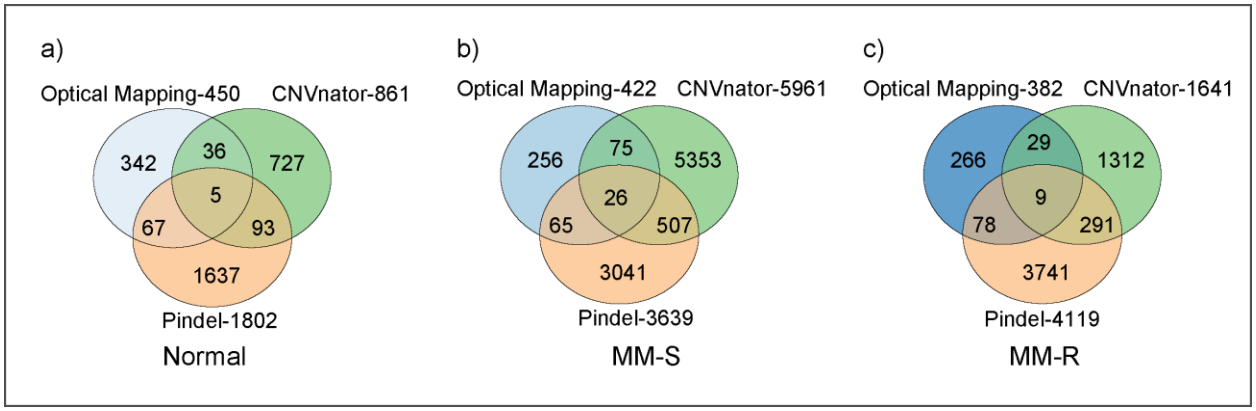

Fig. 3-5: Intersection Venn diagrams of insertion calls from Optical Mapping with duplication calls from CNVnator and long insertion calls from Pindel for Normal (a), MM-S (b) and MM-R (c) samples. The text around the circles labels the approach followed by total number of insertions identified by that approach using filters defined in text.

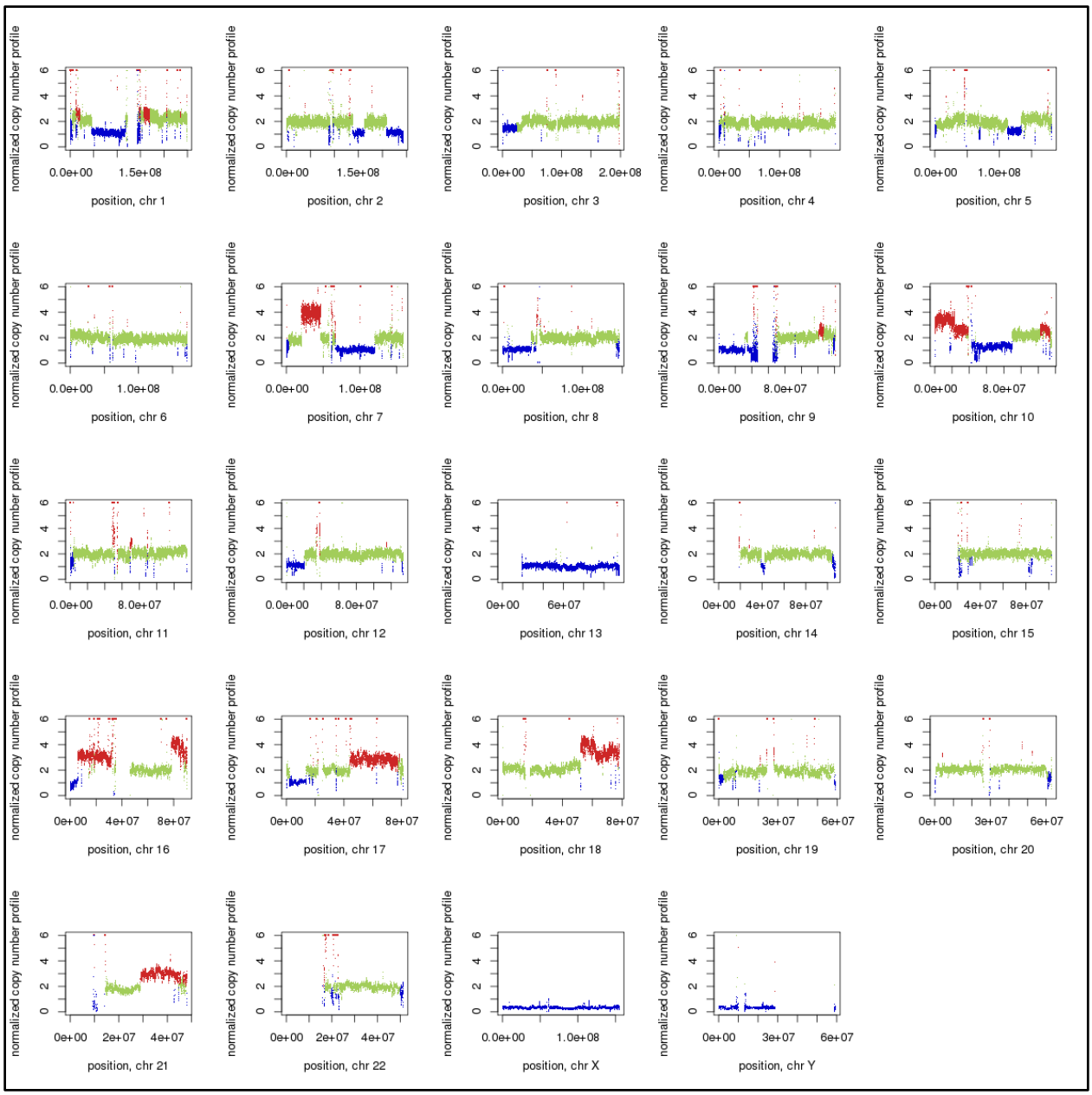

Fig. 3-6: Somatic copy number analysis for MM-S *vs* Normal sample using paired-end DNA sequencing data. Green color indicates normal copy number; red color indicates copy number gain and blue color indicates copy number loss. Chromosomal locations are plotted along X-Axis and normalized copy number profiles for MM-S sample are plotted along Y-Axis.

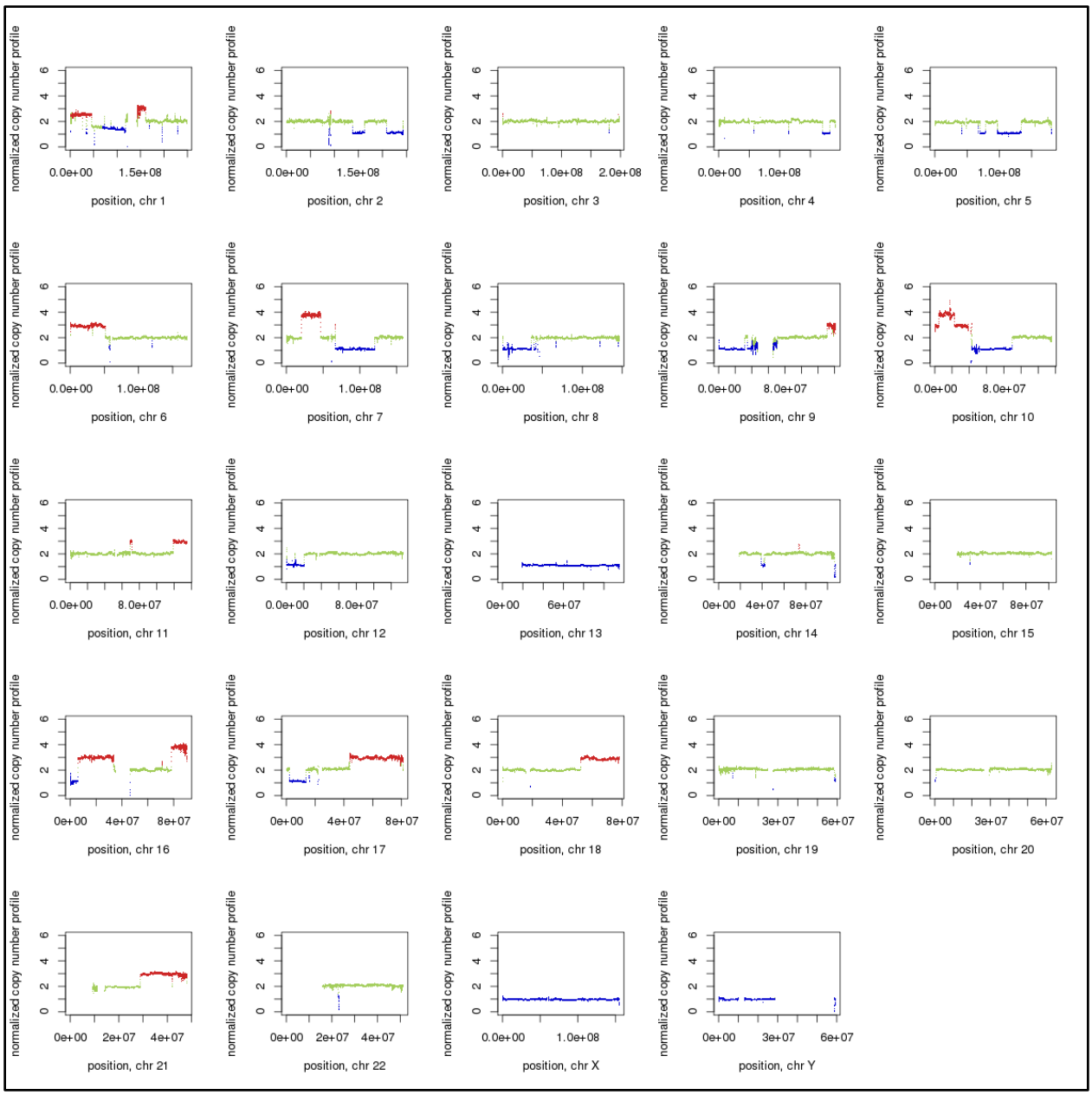

Fig. 3-7: Somatic copy number analysis for MM-R *vs* Normal sample using paired-end DNA sequencing data. Green color indicates normal copy number; red color indicates copy number gain and blue color indicates copy number loss. Chromosomal locations are plotted along X-Axis and normalized copy number profiles for MM-R sample are plotted along Y-Axis.

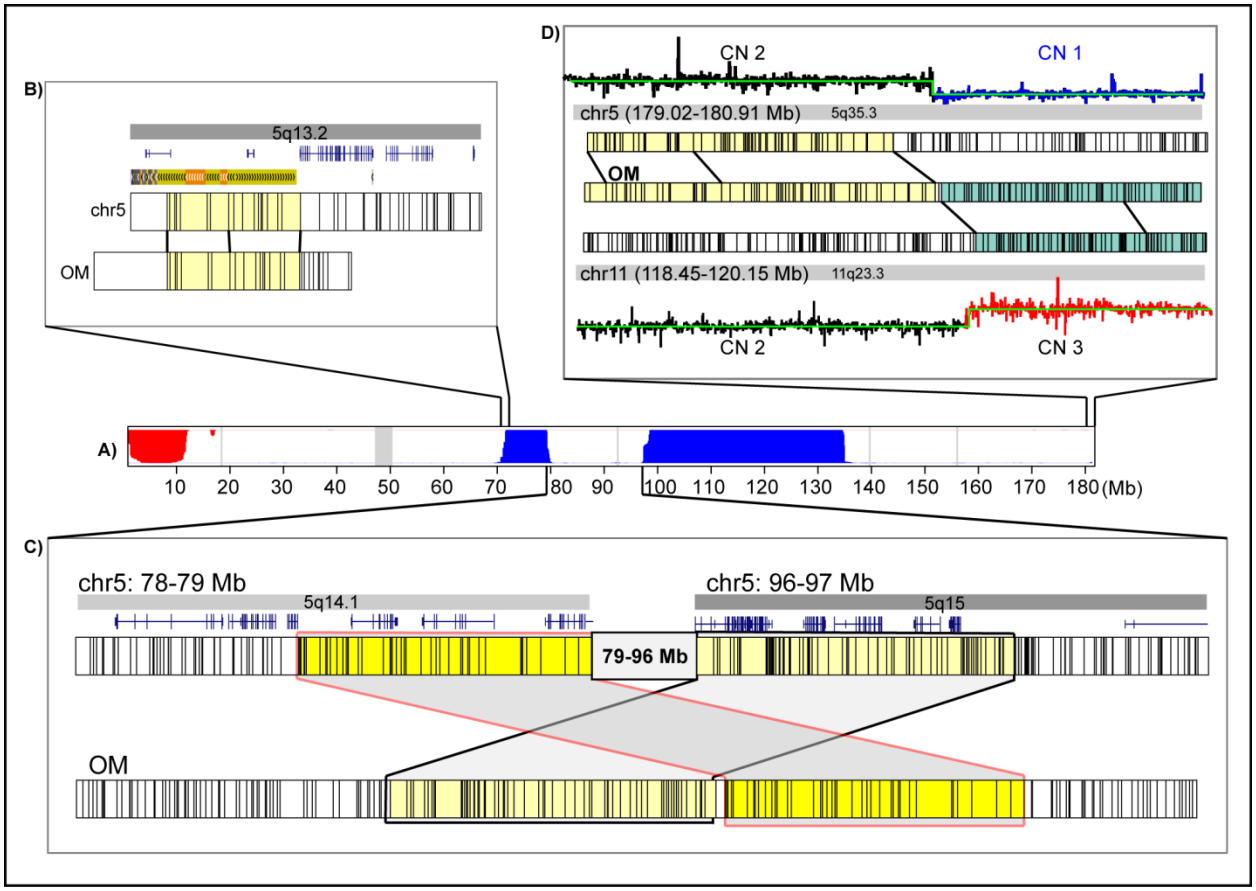

Fig. 3-8: Copy number (CN) alterations and architecture of chr5 in multiple myeloma. (A) Somatic CN analysis of chr5 from MM-R *vs.* Normal Rmap alignments indicates two regions of CN loss (blue): 70.74 Mb - 78.41 Mb and 96.63 Mb - 134.00 Mb. Also, CN neutral regions following 70.74 Mb breakpoint show loss of heterozygosity (LOH), indicating duplication of these regions. B) Chromosomal truncation at 70.74 Mb breakpoint: Consensus optical maps terminate at this breakpoint during assembly, indicating chromosomal truncation. Alignment of consensus map (track: OM) to *in silico* reference map (track: chr5) represents this truncation. *In silico* reference map for chr5 is annotated with chromosomal band, RefSeq genes and segmental duplications (top to bottom) from UCSC genome browser. C) Tandem duplication of 78.40 Mb - 96.63 Mb explains CN neutral LOH: At 96.63 Mb breakpoint, a consensus optical map (track: OM) aligned to two regions on chr5, around 96.63 Mb (alignment with black outline) and around 78.41 Mb (alignment with red outline), indicating tandem duplication of ~18 Mb from 78.40 - 96.63 Mb. We could not resolve the 134.00 Mb breakpoint using Optical Mapping or DNA sequencing data. D) t(5;11)(q35.3;q23.3) translocation at 180 Mb breakpoint on chr5 leads to terminal deletion on chr5 and amplification on chr11: One of the consensus optical maps (track: OM) aligns to chr5 and chr11 (tracks: chr5 and chr11), and indicates a translocation. The topmost and bottommost tracks (indicating CN) are modified from CVNnator output and mark regions of CN loss (CN:1; blue) and gain (CN:3; red) that are involved in this unbalanced translocation.

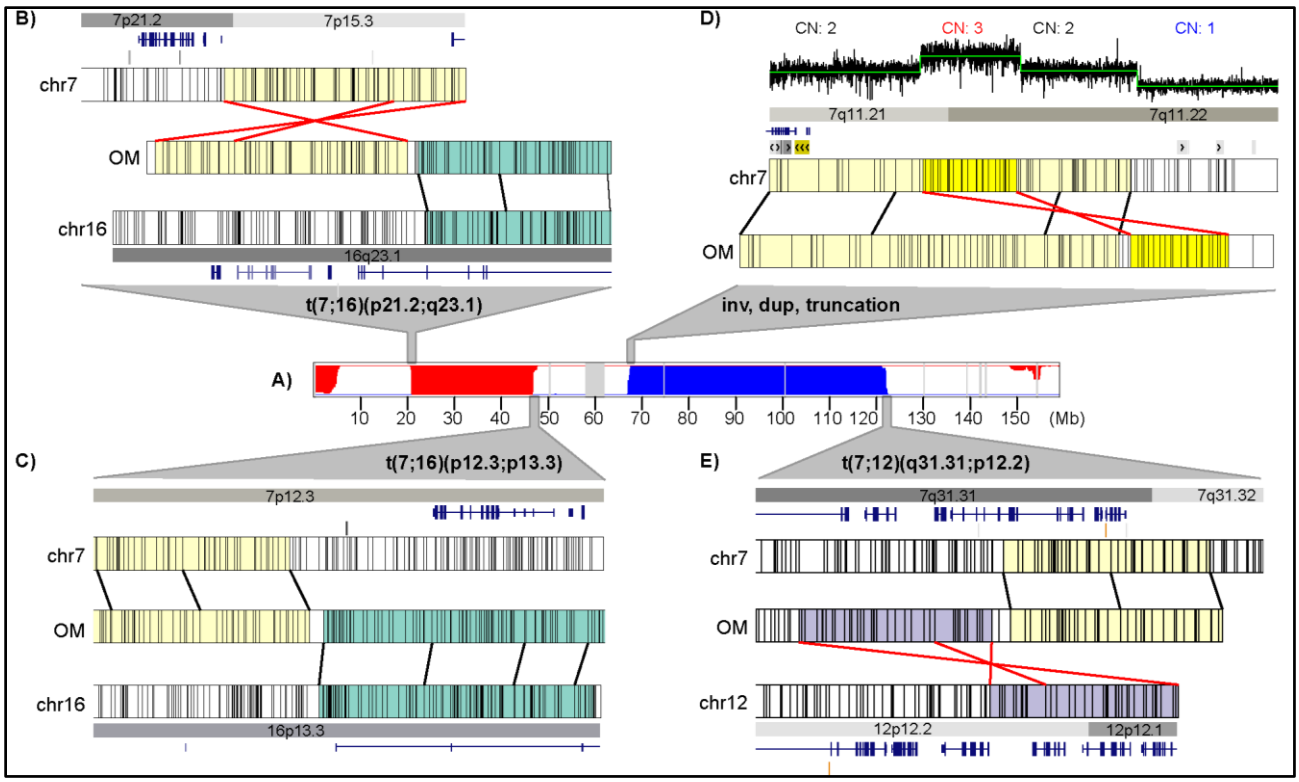

Fig. 3-9: Copy number (CN) alterations and architecture of chr7 in multiple myeloma. A) Somatic CN analysis of chr7 from MM-R *vs.* Normal Rmap alignments indicates a region of CN gain (red; CN:4 from DNA sequencing data) from 20.87 Mb - 46.96 Mb and another region of CN loss (blue; CN:1) from 67.26 Mb - 120.75 Mb. Panels B-E show genomic rearrangements at these CN breakpoints, identified from Rmap iterative assembly. Tracks marked by chr7, chr12 and chr16 represent *in silico* reference maps and are annotated with three tracks: chromosomal band, RefSeq genes and segmental duplications (top to bottom) from UCSC genome browser. Traces marked by OM represent consensus optical maps obtained from Rmap iterative assembly. Connecting lines between reference maps and consensus maps tracks indicate alignments. B) t(7;16)(p21.2;q23.1): A translocation between chr7 (20.87 Mb) and chr16 (78.29 Mb). C) t(7;16)(p12.3;p13.3): Another translocation between chr7 (46.96 Mb) and chr16 (6.00 Mb). Both B and C explain CN gain at these breakpoints on chr7 and chr16. D) An inverse (inv) duplication (dup) leads to truncation of chr7 at 67.2 Mb. The CN plot (shown on top of chromosomal band) is obtained from CNVnator output and shows local CN changes. E) t(7;12)(q31.31;p12.2): A translocation between chr7 and chr12, which is also associated with 12p loss.

3.8. Tables

Table 3-1: Summary statistics of DNA sequencing data collected for Normal, MM-S and MM-R samples.

Parameter	Normal	MM-S	MM-R
Total reads (#)	1,854,980,852	2,465,732,998	3,109,442,418
Properly paired #(% total)	1,749,903,124 (94.34)	2,283,633,732 (92.61)	2,947,025,746 (94.78)
Coverage (X)	56.98	68.41	91.32
Duplication rate (%)	4.66	12.33	9.73
Insert size in bp (mean (sd))	245.37 (72.80)	239.17 (70.02)	255.09 (77.16)

Table 3-2: Summary of GWAS SNPs found in Normal, MM-S and MM-R samples.

SNP ID	Chr	Location*	Normal	MM-S	MM-R	Study	Ref>Sample	Type
rs10936599	3q26.2	41,925,398	Yes	Yes	Yes	Chubb 2013	CC>TT	Homozygous
rs2285803	6p21.33	31,107,258	Yes	Yes	Yes	Chubb 2013	TT>CT	Heterozygous
rs4273077	17p11.2	16,849,139	No	No	No	Chubb 2013	AA>AA	Reference
rs877529	22q13.1	39,542,292	Yes	Yes	Yes	Chubb 2013	GG>AA	Homozygous
rs1052501	3p22.1	41,928,398	Yes	Yes	Yes	Broderick 2012	CC>TT	Homozygous
rs4487645	7p15.3	21,938,240	Yes	Yes	Yes	Broderick 2012	CC>AC	Heterozygous
rs6746082**	2p23.3	25,659,244	Yes	Yes	Yes	Broderick 2012	AA>AC	Heterozygous
rs9344	11q13.3	69,462,890	Yes	Yes	Yes	Weinhold 2013	GG>AG	Heterozygous

*Based on UCSC hg19 coordinates

**Mentioned as a promising association in Broderick et al. 2012

Table 3-3: Summary of single nucleotide variants found in MM-S and MM-R samples.

Sample	Total SNVs	SNVs/Mb	Synonymous (S)	Non-synonymous (NS)*	NS:S
MM-S	10,224	3.529	27	90 (30,60,0)**	3.33
MM-R	13,511	4.523	33	101 (0,60,41)**	3.06

*Includes missense, nonsense and splice-site variants.
**The numbers in parenthesis indicate the number of unique to MM-S sample, shared between MM-S and MM-R samples, and unique to MM-R sample respectively.

Table 3-4: Nonsynonymous SNVs in MM-S sample.

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
1	7890053	G	A/G	missense	1776342	ALA>THR	benign	PER3	Yes
1	95307582	T	C/T	missense	0	TYR>HIS	possibly-damaging	SLC44A3	
1	145588435	C	C/G	missense	0	GLU>ASP	possibly-damaging	NUDT17	
1	158064536	G	A/G	missense	0	ASP>ASN	probably-damaging	KIRREL	Yes
1	160319998	C	C/G	missense	0	ILE>MET	probably-damaging	NCSTN	
1	180145170	T	C/T	splice-donor	0	none	unknown	QSOX1	
2	71825804	T	G/T	missense	0	PHE>VAL	probably-damaging	DYSF	
2	90414180	A	A/T	missense	0	LEU>HIS	unknown	LOC101060017	
2	163137991	G	C/G	missense	0	ILE>MET	probably-damaging	IFIH1	
2	201752337	A	A/T	missense-near-splice	0	MET>LYS	probably-damaging	PPIL3	
2	206592707	C	A/C	stop-gained	0	TYR>stop	unknown	NRP2	
2	220504235	C	C/G	missense	0	LEU>VAL	benign	SLC4A3	
3	36899272	T	G/T	missense	0	LYS>ASN	benign	TRANK1	
3	72842158	C	A/C	missense	0	ASP>TYR	probably-damaging	SHQ1	
3	78763667	T	A/T	missense	0	ILE>PHE	probably-damaging	ROBO1	
3	107491732	T	A/T	missense	0	ASP>GLU	benign	BBX	
4	53610885	G	A/G	missense	0	THR>ILE	probably-damaging	ERVMER34-1	
4	57180974	C	C/G	missense	0	LEU>VAL	benign	KIAA1211	
4	150607557	G	G/T	missense	0	ALA>GLU	unknown	LOC285423	
5	54456949	T	G/T	missense	0	PHE>CYS	probably-damaging	CDC20B,GPX8	
5	58270678	G	A/G	missense	0	THR>MET	probably-damaging	PDE4D	
5	90085612-13	CC	CC/AA	missense	0	PRO>ASN	probably-damaging	GPR98	
5	134015373	G	C/G	missense	0	ASP>HIS	probably-damaging	SEC24A	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
6	29408397	C	C/G	missense	143307828	THR>ARG	benign	OR11A1,OR10C1	
6	38885876	C	C/T	missense	0	ALA>VAL	benign	DNAH8, LOC100131047	
6	40400315	G	G/T	missense	0	LEU>MET	probably-damaging	LRFN2	
6	47574012	C	C/T	missense-near-splice	0	SER>PHE	probably-damaging	CD2AP	
6	135511323	T	G/T	stop-gained	0	LEU>stop	unknown	MYB	
7	1512765	G	G/T	missense	0	LEU>ILE	benign	INTS1	
7	120764440	A	A/G	missense	0	ASP>GLY	possibly-damaging	CPED1	
8	77776292	C	C/T	stop-gained	0	GLN>stop	unknown	ZFHX4	
9	34490479	C	C/T	missense	0	PRO>LEU	benign	DNAI1	
10	15714671	C	C/G	missense	0	GLU>GLN	probably-damaging	ITGA8	
10	28878776	G	G/T	stop-gained	77910453	GLU>stop	unknown	WAC	
10	74103246	T	C/T	missense	0	TYR>CYS	probably-damaging	DNAJB12	
11	3381109	C	C/T	missense	0	ALA>THR	probably-damaging	ZNF195	
11	55110990	T	G/T	missense	0	LEU>ARG	possibly-damaging	OR4A16	
11	55761855	C	A/C	missense	0	ALA>SER	benign	OR5F1	
11	56127807	C	A/C	missense	0	LEU>MET	possibly-damaging	OR8J1	
11	58978858	G	A/G	missense	0	ALA>VAL	probably-damaging	MPEG1	
11	64662883	G	A/G	missense	0	SER>PHE	probably-damaging	ATG2A	
11	64803001	C	C/T	missense	142074663	PRO>LEU	benign	SNX15, ARL2-SNX15	
11	76814330	G	G/T	missense	0	VAL>LEU	benign	CAPN5,OMP	
11	76893150	C	C/T	missense	376121363	ARG>TRP	probably-damaging	MYO7A	
11	119228018	C	C/G	splice-acceptor	0	none	unknown	USP2	
12	32977013	C	C/G	missense	0	ARG>THR	probably-damaging	PKP2	
12	49444852	C	C/T	missense	0	GLY>SER	benign	KMT2D	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
12	65224215	C	C/T	missense	0	THR>ILE	possibly-damaging	TBC1D30	
12	89917265	G	C/G	missense	0	ILE>MET	probably-damaging	GALNT4,POC1B,POC1B-GALNT4	
12	101336194	C	C/T	stop-gained	0	ARG>stop	unknown	ANO4	
12	112165766	T	G/T	missense-near-splice	0	SER>ARG	benign	ACAD10	
13	25671027	A	A/G	missense	78826513	LYS>GLU	probably-damaging	PABPC3	
13	32937348	C	A/C	stop-gained	80359035	SER>stop	unknown	BRCA2	
13	58299148	C	C/G	missense	0	ALA>GLY	benign	PCDH17	
14	36155788	T	C/T	missense	0	ASP>GLY	possibly-damaging	RALGAPA1	
14	65544630	C	C/T	splice-donor	0	none	unknown	MAX	
14	105405907	G	G/T	missense	0	THR>ASN	possibly-damaging	AHNAK2	
14	105517579	C	A/C	missense	0	VAL>LEU	benign	GPR132	
15	34543100	C	A/C	missense-near-splice	0	GLY>CYS	probably-damaging	SLC12A6	
15	38243253	T	A/T	missense	0	PHE>ILE	possibly-damaging	TMCO5A	
15	49254992	A	A/G	missense	0	LEU>SER	benign	SHC4	
15	63937706	G	A/G	missense	0	SER>PHE	benign	HERC1	
15	65916544	G	C/G	missense	0	ARG>SER	benign	SLC24A1	
16	720261	C	C/G	missense	0	ILE>MET	probably-damaging	RHOT2	
16	29917181	G	A/G	missense	0	ARG>HIS	probably-damaging	ASPHD1	
16	50332935	G	G/T	missense	0	ALA>SER	possibly-damaging	ADCY7	
16	57054875	T	A/T	missense	0	LEU>GLN	probably-damaging	NLRC5	
16	66943991	G	G/T	missense	0	HIS>ASN	probably-damaging	CDH16	
17	3985740	G	C/G	missense	185558294	THR>ARG	possibly-damaging	ZZEF1	
17	27065255	G	G/T	missense	0	CYS>PHE	probably-damaging	NEK8	
17	33328991	T	C/T	missense	0	SER>PRO	benign	LIG3	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
17	47236571	T	G/T	missense	0	VAL>GLY	probably-damaging	B4GALNT2	
17	78320159	A	A/G	missense	0	HIS>ARG	benign	RNF213	
18	7050694	T	G/T	missense-near-splice	0	GLU>ALA	probably-damaging	LAMA1	
18	63511049	A	A/T	missense-near-splice	0	GLU>VAL	possibly-damaging	CDH7	
19	8953362	A	A/T	missense	0	LYS>MET	probably-damaging	MBD3L1	
19	10088132	G	A/G	missense	0	PRO>LEU	probably-damaging	COL5A3	
19	12125727	C	A/C	missense	0	GLY>VAL	possibly-damaging	ZNF433	
19	37854430	C	A/C	missense	0	ALA>ASP	possibly-damaging	HKR1	
19	53959409	T	A/T	missense	0	TYR>ASN	probably-damaging	ZNF761	
19	55178201	G	G/T	splice-donor	199892280	none	unknown	LILRB4	Yes
20	24930200	G	G/T	missense	0	GLY>VAL	benign	CST7	
20	48160968	T	A/T	missense	0	GLU>VAL	benign	PTGIS	
21	39772519	G	A/G	missense	0	THR>MET	probably-damaging	ERG	
21	41032924-25	TC	TC/AT	stop-gained	0	CYS>stop	unknown	B3GALT5	
22	32111069	T	C/T	missense	0	TYR>CYS	benign	PRR14L	
22	37581309	G	G/T	missense	0	PRO>THR	benign	C1QTNF6	
22	41572348	G	G/T	missense	0	GLY>VAL	probably-damaging	EP300	

Table 3-5: Nonsynonymous SNVs in MM-R sample.

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
1	95307582	T	C/T	missense	0	TYR>HIS	possibly-damaging	SLC44A3	
1	145588435	C	C/G	missense	0	GLU>ASP	possibly-damaging	NUDT17	
1	154034081	G	A/G	missense	0	SER>LEU	benign	NUP210L	
1	158064536	G	A/G	missense	0	ASP>ASN	probably-damaging	KIRREL	Yes
1	171482265	G	C/G	missense	376186276	ASP>HIS	probably-damaging	PRRC2C	
1	182872205	G	C/G	stop-gained	0	SER>stop	unknown	SHCBP1L	
2	37873214	C	A/C	missense	0	GLY>CYS	possibly-damaging	CDC42EP3	
2	97031715	A	A/T	missense	0	GLN>HIS	benign	NCAPH	
2	122139861	G	C/G	missense	0	HIS>GLN	benign	CLASP1	
2	131984477	T	A/T	missense	0	LEU>ILE	benign	POTEE	
2	179636128	T	A/T	missense	0	GLU>ASP	probably-damaging	TTN	
2	201752337	A	A/T	missense-near-splice	0	MET>LYS	probably-damaging	PPIL3	
2	203407058	T	G/T	missense	0	MET>ARG	possibly-damaging	BMPR2	
2	220504235	C	C/G	missense	0	LEU>VAL	benign	SLC4A3	
3	36899272	T	G/T	missense	0	LYS>ASN	benign	TRANK1	
3	72842158	C	A/C	missense	0	ASP>TYR	probably-damaging	SHQ1	
3	78763667	T	A/T	missense	0	ILE>PHE	probably-damaging	ROBO1	
3	100569525	C	C/T	missense	0	GLU>LYS	probably-damaging	ABI3BP	
3	107491732	T	A/T	missense	0	ASP>GLU	benign	BBX	
3	183907402	G	A/G	missense	0	ALA>THR	benign	ABCF3	
3	195508067	T	G/T	missense	0	THR>PRO	benign	MUC4	
4	53610885	G	A/G	missense	0	THR>ILE	probably-damaging	ERVMER34-1	
4	57180974	C	C/G	missense	0	LEU>VAL	benign	KIAA1211	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
4	123183994	G	A/G	missense	0	ALA>THR	benign	KIAA1109	
4	150607557	G	G/T	missense	0	ALA>GLU	unknown	LOC285423	
4	176561315	C	C/T	missense	0	ALA>THR	benign	GPM6A	
5	54456949	T	G/T	missense	0	PHE>CYS	probably-damaging	CDC20B, GPX8	
5	58270678	G	A/G	missense	0	THR>MET	probably-damaging	PDE4D	
5	134015373	G	C/G	missense	0	ASP>HIS	probably-damaging	SEC24A	
5	142421402	A	A/T	missense	0	TYR>PHE	probably-damaging	ARHGAP26	
6	29408397	C	C/G	missense	143307828	THR>ARG	benign	OR11A1, OR10C1	
6	38885876	C	C/T	missense	0	ALA>VAL	benign	DNAH8, LOC100131047	
6	40400315	G	G/T	missense	0	LEU>MET	probably-damaging	LRFN2	
6	47574012	C	C/T	missense-near-splice	0	SER>PHE	probably-damaging	CD2AP	
6	71442079	A	A/C	splice-acceptor	0	none	unknown	SMAP1	
6	125116020	T	C/T	missense	0	ILE>THR	unknown	NKAIN2	
7	1512765	G	G/T	missense	0	LEU>ILE	benign	INTS1	
7	82545934	C	C/T	missense	0	GLU>LYS	probably-damaging	PCLO	
7	120764440	A	A/G	missense	0	ASP>GLY	possibly-damaging	CPED1	
8	27516342	G	G/T	missense	0	ALA>SER	benign	SCARA3	
8	77776292	C	C/T	stop-gained	0	GLN>stop	unknown	ZFHX4	
8	103324003	C	A/C	missense	0	GLY>VAL	probably-damaging	UBR5	
8	139163948	A	A/G	missense	0	SER>PRO	possibly-damaging	FAM135B	
8	141874481	C	A/C	missense	0	ARG>LEU	probably-damaging	PTK2	
9	23692750-51	CA	CA/AT	missense	0	MET>ASN	benign	ELAVL2	
9	32418440	T	G/T	missense	0	TYR>ASP	probably-damaging	ACO1	
9	78953349	G	C/G	missense	0	CYS>SER	unknown	PCSK5	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
10	18254587	C	C/T	missense	0	SER>PHE	benign	SLC39A12	
10	64967054	C	A/C	missense	0	ALA>SER	benign	JMJD1C	
10	74103246	T	C/T	missense	0	TYR>CYS	probably-damaging	DNAJB12	
11	58978858	G	A/G	missense	0	ALA>VAL	probably-damaging	MPEG1	
11	64662883	G	A/G	missense	0	SER>PHE	probably-damaging	ATG2A	
11	64803001	C	C/T	missense	142074663	PRO>LEU	benign	SNX15, ARL2-SNX15	
11	67264780	C	A/C	missense	0	VAL>PHE	probably-damaging	PITPNM1	
11	76814330	G	G/T	missense	0	VAL>LEU	benign	CAPN5,OMP	
11	76893150	C	C/T	missense	376121363	ARG>TRP	probably-damaging	MYO7A	
11	111784473	G	A/G	missense	0	ALA>THR	benign	HSPB2,HSPB2-C11orf52	
11	119228018	C	C/G	splice-acceptor	0	none	unknown	USP2	
12	14804446	C	C/T	splice-acceptor	0	none	unknown	GUCY2C	
12	32977013	C	C/G	missense	0	ARG>THR	probably-damaging	PKP2	
12	39154675	C	C/G	missense	0	GLU>ASP	benign	CPNE8	
12	49446851	C	C/T	missense	0	CYS>TYR	probably-damaging	KMT2D	
12	49496304	T	A/T	missense	0	GLU>VAL	probably-damaging	LMBR1L	
12	81004359	G	G/T	stop-gained	0	GLU>stop	unknown	PTPRQ	
12	89917265	G	C/G	missense	0	ILE>MET	probably-damaging	GALNT4,POC1B,POC1B-GALNT4	
12	101336194	C	C/T	stop-gained	0	ARG>stop	unknown	ANO4	
13	25671027	A	A/G	missense	78826513	LYS>GLU	probably-damaging	PABPC3	
13	58299148	C	C/G	missense	0	ALA>GLY	benign	PCDH17	
14	36155788	T	C/T	missense	0	ASP>GLY	possibly-damaging	RALGAPA1	
14	65544630	C	C/T	splice-donor	0	none	unknown	MAX	
14	75329621	G	G/T	missense	0	ALA>ASP	possibly-damaging	PROX2	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
14	76156638	C	A/C	missense	0	LEU>MET	possibly-damaging	TTLL5	
14	105405907	G	G/T	missense	0	THR>ASN	possibly-damaging	AHNAK2	
14	105517579	C	A/C	missense	0	VAL>LEU	benign	GPR132	
15	38243253	T	A/T	missense	0	PHE>ILE	possibly-damaging	TMCO5A	
15	63937706	G	A/G	missense	0	SER>PHE	benign	HERC1	
15	65916544	G	C/G	missense	0	ARG>SER	benign	SLC24A1	
16	720261	C	C/G	missense	0	ILE>MET	probably-damaging	RHOT2	
16	31539933	T	C/T	missense	0	LEU>PRO	probably-damaging	AHSP	
16	57054875	T	A/T	missense	0	LEU>GLN	probably-damaging	NLRC5	
16	66943991	G	G/T	missense	0	HIS>ASN	probably-damaging	CDH16	
17	33328991	T	C/T	missense	0	SER>PRO	benign	LIG3	
17	47236571	T	G/T	missense	0	VAL>GLY	probably-damaging	B4GALNT2	
17	78320159	A	A/G	missense	0	HIS>ARG	benign	RNF213	
18	7050694	T	G/T	missense-near-splice	0	GLU>ALA	probably-damaging	LAMA1	
18	63511049	A	A/T	missense-near-splice	0	GLU>VAL	possibly-damaging	CDH7	
19	8953362	A	A/T	missense	0	LYS>MET	probably-damaging	MBD3L1	
19	10088132	G	A/G	missense	0	PRO>LEU	probably-damaging	COL5A3	
19	10822849	G	A/G	missense	0	ARG>GLN	probably-damaging	QTRT1	
19	37854430	C	A/C	missense	0	ALA>ASP	possibly-damaging	HKR1	
19	52004585	C	C/T	missense	0	ASP>ASN	benign	SIGLEC12	
19	55178201	G	G/T	splice-donor	199892280	none	unknown	LILRB4	Yes
20	48160968	T	A/T	missense	0	GLU>VAL	benign	PTGIS	
21	39772519	G	A/G	missense	0	THR>MET	probably-damaging	ERG	
21	41032924-25	TC	TC/AT	stop-gained	0	CYS>stop	unknown	B3GALT5	

chr	position	referenceBase	sampleAlleles	functionGVS	rsID	aminoAcids	polyPhen	geneList	COSMIC
22	32111069	T	C/T	missense	0	TYR>CYS	benign	PRR14L	
22	37581309	G	G/T	missense	0	PRO>THR	benign	C1QTNF6	
22	41572348	G	G/T	missense	0	GLY>VAL	probably-damaging	EP300	
22	41922443	G	G/T	missense	0	ALA>SER	probably-damaging	ACO2,POLR3H	

Table 3-6: Regions of copy number neutral loss of heterozygosity in MM-R sample.

Chr	Start	End
1	160,523,100	249,250,621
5	78,413,500	96,631,201
5	134,007,800	180,039,800
14	42,142,500	106,107,000

Table 3-7: Summary of somatic deletions found in MM-S and MM-R samples.

Size (bp)	Count	Overlapping genes (#)	Overlapping exons (#)	Genes affected
Shared between MM-S and MM-R				
10-400*	228	76	6	UBIAD1, GSTA7P, TCL1A, TP53, ASXL3, MTMR1
>400**	38	19	10	CDKN2C/FAF1, DPP10, NOSTRIN, CRYBG3, IGF2BP2, SLC4A4, CAMK2D, SORBS2, ADAMTS6, PIK3R1, CPED1, GRM8, ADRB3, E2F8, SCFD1, KLC1, FSD2, ELL, CSNK2A1
Novel to MM-R only				
10-400*	129	45	2	CCNG2, MAPT
>400**	27	13	7	GABRD, MIR1976/RPS6KA1, CTNNA2, PTPN4, FANCD20S/FANCD2, MRPS36, VCPIP1/MYBL1, MALRD1, SBF2, MIR3922/CHST11, ZWILCH, NFIX, DSCAM

*For deletions 10-400 bp in size, only the ones overlapping exons have been listed.
**For deletions >400 bp in size, the ones overlapping genes have been listed.

Table 3-8: Somatic deletions shared between MM-S and MM-R samples.

Chr	Left	Right	Size (bp)	GeneName	Exon
*10 – 400 bp**					
1	11,348,152	11,348,178	24	UBIAD1	Yes
6	52,609,304	52,609,348	41	GSTA7P	Yes
14	96,176,620	96,176,683	61	TCL1A	Yes
17	7,577,580	7,577,656	73	TP53	Yes
18	31,319,579	31,319,689	105	ASXL3	Yes
X	149,905,913	149,905,947	30	MTMR1	Yes
*Larger than 400 bp***					
1	51,388,063	51,500,646	112,580	CDKN2C, FAF1	Yes
2	115,852,008	115,852,513	497	DPP10	
2	169,716,286	169,718,523	2,232	NOSTRIN	Yes
3	97,592,838	97,602,844	10,005	CRYBG3	Yes
3	185,382,002	185,384,817	2,811	IGF2BP2	
4	72,055,755	72,058,776	3,017	SLC4A4	
4	114,366,543	114,560,786	194,238	CAMK2D	Yes
4	186,712,618	186,713,167	548	SORBS2	
5	64,625,196	64,626,109	909	ADAMTS6	Yes
5	67,521,618	67,591,647	70,026	PIK3R1	Yes
7	120,758,274	120,759,200	923	CPED1	
7	126,565,706	126,568,527	2,820	GRM8	
8	37,815,241	37,825,152	9,906	ADRB3	Yes
11	19,246,933	19,261,221	14,286	E2F8	Yes
14	31,167,244	31,169,371	2,126	SCFD1	
14	104,112,111	104,135,418	23,304	KLC1	Yes
15	83,443,876	83,446,124	2,245	FSD2	
19	18,584,289	18,617,985	33,692	ELL	
20	498,237	514,560	16,322	CSNK2A1	Yes

*For deletions between 10-400 bp in size, only the ones overlapping exons have been listed.
**For deletions larger than 400 bp, only the ones overlapping genes have been listed.

Table 3-9: Somatic deletions novel to the MM-R sample.

Chr	Left	Right	Size (bp)	Gene	GeneName	Exon
*10 - 400 bp**						
4	78079672	78079731	55	Yes	CCNG2	Yes
17	44060879	44060918	38	Yes	MAPT	Yes
*Larger than 400 bp***						
1	1,961,642	1,966,941	5,294	Yes	GABRD	Yes
1	26,878,599	26,920,955	42,352	Yes	MIR1976, RPS6KA1	Yes
2	79,949,929	79,953,973	4,042	Yes	CTNNA2	
2	120,524,004	120,537,350	13,342	Yes	PTPN4	
3	10,120,856	10,132,860	12,001	Yes	FANCD20S, FANCD2	Yes
5	68,522,864	68,525,992	3,123	Yes	MRPS36	Yes
8	67,507,450	67,548,757	41,304	Yes	VCPIP1, MYBL1	Yes
10	19,886,704	19,891,496	4,791	Yes	MALRD1	
11	10,034,570	10,049,337	14,764	Yes	SBF2	
12	104,982,273	104,988,306	6,029	Yes	MIR3922, CHST11	Yes
15	66,817,238	66,819,521	2,282	Yes	ZWILCH	
19	13,141,917	13,145,399	3,477	Yes	NFIX	
21	42,070,774	42,221,907	151,127	Yes	DSCAM	Yes

*For deletions between 10-400 bp in size, only the ones overlapping exons have been listed.
**For deletions larger than 400 bp, only the ones overlapping genes have been listed.

Table 3-10: Structural variants that underlie large scale copy number variation in MM-R sample.

Chr1	Loc1	Gene_Loc1	Chr2	Loc2	Gene_Loc2	Event
Chr1	46,360,804	MAST2	Chr1	117,273,548		Deletion
Chr1	160,523,168	CD84	Chr6	51,118,469		Translocation
Chr1	180,927,627		Chr9	131,339,039	SPTAN1	Translocation
Chr2	138,855,727		Chr2	162,035,600	TANK	Deletion
Chr2	208,070,175		Chr10	23,276,740	ARMC3	Translocation
Chr4	170,232,951		Chr4	182,336,593		Deletion
Chr5	70,740,000					Truncation
Chr5	78,413,500	BHMT	Chr5	96,631,200		Tandem Duplication
Chr5	134,007,729	SEC24A	Chr7	66,958,972		Translocation
Chr5	180,039,850	FLT4	Chr11	119,453,437		Translocation
Chr7	20,878,695		Chr16	78,290,600	WWOX	Translocation
Chr7	46,962,661		Chr16	5,998,555		Translocation
Chr7	67,099,767		Chr7	67,262,743		Inverse Duplication
Chr7	120,756,417	CPED1	Chr12	20,883,779	SLCO1C1	Translocation
Chr8	36,284,561		Chr9	32,013,790		Translocation
Chr8	39,393,625		Chr8	41,441,031	AGPAT6	Deletion
Chr11	69,453,552		Chr14	106,107,045		Translocation
Chr14	39,296,961	LINC00639	Chr21	29,006,964	MIR5009	Translocation
Chr14	48,211,973		Chr14	52,920,149	TXNDC16	Inversion
Chr15	30,873,871	ULK4P1 ULK4P2	Chr15	31,189,559		Deletion
Chr17	2,023,044	SMG6	Chr17	13,892,185		Deletion
Chr17	13,892,185		Chr17	15,895,185	ZSWIM7	Inversion
Chr17	15,895,213	ZSWIM7	Chr17	16,236,159		Deletion
Chr17	44,041,355	MAPT	Chr19	58,412,876		Translocation
Chr18	52,209,201		Chr9	Past End		Translocation

Chapter 4: Closing remarks and future directions

In this thesis, we presented an integrative approach that uses Optical Mapping (Schwartz et al. 1993; Dimalanta et al. 2004; Teague et al. 2010) and next generation DNA sequencing (Turcatti et al. 2008; Bentley et al. 2008) to study somatic variation in a tumor genome, with primary focus on structural variation, copy number variation and genomic rearrangements. Through our analysis, we identified and characterized a large number of these variants in a multiple myeloma patient genome. Additionally, we demonstrated that mutational complexity increased with tumor progression across the entire length spectrum of variation. This was indicated by an increase in number of single nucleotide variants, structural variants, copy number variants and genomic rearrangements through our analysis of serially obtained drug sensitive (MM-S) and drug refractory (MM-R) tumor samples from the same patient.

While single nucleotide variants have been assayed in large cohorts of multiple myeloma patients in previous studies (Chapman et al. 2011; Lohr et al. 2014), a detailed analysis of structural variation in multiple myeloma is yet to be reported. Our work has addressed this gap and identified widespread structural variation in multiple myeloma, which is distributed over coding and non-coding regions of the genome. Variants discovered in coding regions of the genome could serve as initial targets for further functional analysis to ascertain their role in multiple myeloma pathogenesis and drug resistance mechanisms. Additionally, it will be advantageous to understand the functional consequences of structural variation and genomic rearrangements that have been discovered in non-coding portion of this tumor genome, based on the recent work by ENCODE consortium, which has shown that a much larger part of the non-coding genome plays an important role in gene regulation than

previously appreciated (Dunham et al. 2012). To refine functional hypothesis, multiplexing to study many analytes (DNA, RNA and proteins) in single experiments can be used to analyze these mutations and their downstream products across a large number of patient samples quickly and cost-effectively (Gerdes et al. 2013). Finally, recently developed genome editing CRISPR-Cas systems (Jinek et al. 2012; Cong et al. 2013) can then be employed to study the effects of these genomic alterations *ex vivo*.

However, our work has two major limitations. First, we are limited by sample size (n=1) and hence, these findings cannot be generalized to multiple myeloma. Analysis of larger cohorts of multiple myeloma patients for structural variation might lead us to a better understanding of patterns of structural variation in multiple myeloma. Additionally, analysis of larger cohorts might help us identify recurring variants, which would be interesting targets for functional validation. Second, our work does not account for intra-tumor heterogeneity, which has recently been revealed for multiple myeloma (Lohr et al. 2014; Bolli et al. 2014) and other cancers (Ding et al. 2012). This issue arises because of the current structural variation identification pipeline, which considers only single representation of the genome (Valouev et al. 2006; Teague et al. 2010). However, multiple genotypes are assembled during iterative assembly and can be identified, but using manual, rather than automated methods.

We believe that collection of deeper and better Rmap datasets can address both of these issues. Increased turnaround times will allow us to analyze samples faster, thereby potentiating analysis of many patient samples. Also, development of experimental pipelines that produce deeper and better data might allow us to do clonal analysis of tumor samples. Our lab has already made substantial progress in achieving these goals. Development of a new

physical mapping system called Nanocoding (Jo et al. 2007; Jo et al. 2009; Kounovsky-Shafer et al. 2013) has potentiated new forms of genome analysis through substantially improved single molecule datasets and enhanced throughput. As discussed in chapter 1, preliminary results from Nanocoding highlight a highly improved route to genome analysis *via* single molecules with the ability to resolve structural variants as small as 1 kb in size and perhaps, even smaller. We have already collected Nmap datasets for two tumor samples and are currently analyzing these genomes. On the computational front, previous work by Dr. Brian Teague, another graduate student from our lab, describes a new Hidden Markov Model (HMM) based computational pipeline for refinement of optical map assemblies (Teague 2012). This pipeline can be used for haplotyping and to study intra-tumor heterogeneity.

In the near future, long sequencing reads generated *via* nanopore sequencing (Branton et al. 2008), single-molecule sequencing using zero mode waveguides (Levene et al. 2003; Eid et al. 2009; Korlach et al. 2010; Huddleston et al. 2014) or novel long-read DNA sequencing technologies might enable unambiguous reconstruction of genomic rearrangements. Until then, Optical Mapping and its successors continue to be a unique system capable of analyzing long-range genome structures, one molecule at a time.

Almost all tumor genomes evolve differently, and consequently present with varied routes to tumor pathogenesis and drug treatments. In near future, development of single-cell based genomic analysis approaches that can comprehensively capture somatic variation in individual tumor cells over time will help us understand these manifestations as functions of underlying genomic changes in highly heterogeneous tumor cell populations. Recent work by Navin and colleagues, which describes single-cell analysis of copy number changes (Navin et al.

2011) and somatic mutations (Navin et al. 2011; Wang et al. 2014) in breast cancer to study tumor evolution using next generation DNA sequencing, is an important step in that direction. The information gleaned from such experiments could then be used to guide genome based cancer therapies. From a clinical perspective, integration of genomic data with epigenomic (Jones & Baylin 2007; Berdasco & Esteller 2010), transcriptomic (Lappalainen et al. 2013), proteomic (Kim et al. 2014), metabolomic (Wishart et al. 2007) and three dimensional chromatin organization (Lieberman-Aiden et al. 2009; Sexton et al. 2012) data in a clinically actionable timeframe promises to dramatically improve and personalize cancer therapies in the near future.

4.1. Bibliography

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–9.

Berdasco, M. & Esteller, M., 2010. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Developmental cell*, 19(5), pp.698–711.

Bolli, N. et al., 2014. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nature communications*, 5, p.2997.

Branton, D. et al., 2008. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10), pp.1146–53.

Chapman, M.A. et al., 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), pp.467–72.

Cong, L. et al., 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, 339(6121), pp.819–23.

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

Ding, L. et al., 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382), pp.506–10.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910), pp.133–8.

Gerdes, M.J. et al., 2013. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29), pp.11982–7.

Huddleston, J. et al., 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research*, 24(4), pp.688–96.

Jinek, M. et al., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, 337(6096), pp.816–21.

Jo, K. et al., 2007. 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), pp.2673–8.

Jo, K., Schramm, T.M. & Schwartz, D.C., 2009. A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding. *Methods in molecular biology (Clifton, N.J.)*, 544, pp.29–42.

Jones, P.A. & Baylin, S.B., 2007. The epigenomics of cancer. *Cell*, 128(4), pp.683–92.

Kim, M.-S. et al., 2014. A draft map of the human proteome. *Nature*, 509(7502), pp.575–81.

Korlach, J. et al., 2010. Real-time DNA sequencing from single polymerase molecules. *Methods in enzymology*, 472, pp.431–55.

Kounovsky-Shafer, K.L. et al., 2013. Presentation of large DNA molecules for analysis as nanoconfined dumbbells. *Macromolecules*, 46(20), pp.8356–8368.

Lappalainen, T. et al., 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468), pp.506–11.

Levene, M.J. et al., 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, N.Y.)*, 299(5607), pp.682–6.

Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950), pp.289–93.

Lohr, J.G. et al., 2014. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell*, 25(1), pp.91–101.

Navin, N. et al., 2011. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341), pp.90–4.

Schwartz, D.C. et al., 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science (New York, N.Y.)*, 262(5130), pp.110–4.

Sexton, T. et al., 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3), pp.458–72.

Teague, B. et al., 2010. High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 107(24), pp.10848–53.

Teague, B.P., 2012. Experimental and computational advances for studying the human genome with optical mapping. PhD thesis, University of Wisconsin, Madison.

Turcatti, G. et al., 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4), p.e25.

Valouev, A. et al., 2006. 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), pp.15770–5.

Wang, Y. et al., 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513), pp.155–60.

Wishart, D.S. et al., 2007. HMDB: the Human Metabolome Database. *Nucleic acids research*, 35(Database issue), pp.D521–6.