PhD-FSTC-2015-50
The Faculty of Sciences

DISSERTATION

Presented on 03/11/2015 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Sarah KILLCOYNE
Born on 21 January 1978 in Colorado (USA)

INSILICO GENOMES FOR HIGH-THROUGHPUT SEQUENCING CANCER-SPECIFIC ANALYSIS

Dissertation defense committee

Dr Antonio DEL SOL MESA, dissertation supervisor
Assoc. Professor, Université du Luxembourg

Dr David GALAS
Pacific Northwest Diabetes Research Institute
Seattle, USA

Dr Reinhard SCHNEIDER, Chairman
Professor, Université du Luxembourg

Dr Miguel ANDRADE
Assoc. Professor, Max Delbrück Center for Molecular Medicine
Berlin, Germany

Dr Rudi BALLING, Vice Chairman
Professor, Université du Luxembourg

Dissertation by

Sarah Killcoyne

Luxembourg Centre for Systems Biomedicine,

University of Luxembourg

2015

Insilico genomes for high-throughput sequencing cancer-specific analysis

Supervisor: Associate Prof. Antonio del Sol

Dissertation Defense Committee:

Dr. Reinhard Schneider – LCSB, Luxembourg (Chair)

Dr. David Galas – PNDRI, Seattle

Dr. Miguel Andrade – MDC, Germany

Associate Prof. Antonio del Sol – LCSB, Luxembourg

Prof. Rudi Balling – LCSB, Luxembourg

AFFIDAVIT

I hereby confirm that the PhD thesis entitled “Insilico genomes for high-throughput sequencing cancer-specific analysis” has been written independently and without any other sources than cited.

Sarah Killcoyne

Luxembourg

November 12, 2015

ACKNOWLEDGEMENTS

No PhD is done entirely alone and I have many people and institutions to thank for supporting me over the last 4 years. First and foremost I need to thank the LCSB for providing a great environment. It is truly a unique and special place that has been created by the people working here.

For supporting my work and creating a team where everyone could ask questions and propose ideas I need to thank Dr. Antonio del Sol. His encouragement and support allowed me to aim for bigger, more interesting ideas and achievements. My productivity was limited only by my capacity to take in new ideas, and develop in new directions. I would also like to thank my evaluation committee Dr. David Galas and Dr. Rudi Balling. Their support, observations, and suggestions were invaluable in focusing my research and ultimately my thesis.

At the LCSB I have many other people to thank for talking through ideas no matter how poorly formed they were. Dr Gökhan Ertaylan helped a great deal with both his knowledge of cancer pathways and genomics, and his willingness to work through new ideas and directions with me. Abhimanyu Krishna, Susanne Reinsbach, Dr. Thanneer Malai Perumal, Dr Isaac Crespo also provided great feedback on ideas or problems I had whenever I needed it. Thanks as well to Kavitha Rege who took a big risk in doing her MSc work with me, and whose ideas and work will be useful to me in future projects. I also need to thank two people in particular at the Sanger Institute for their willingness to discuss my ideas and answer questions. Drs. David Wedge and Yilong Li in the Cancer Genome Project.

At a practical level I need to thank the Fonds Nationale de la Recherche Luxembourg for providing me with the grant funding necessary to complete my work, Amazon Research & Education grants for access to the AWS infrastructure, the TCGA Research Network for providing the data, and the HPC team at the University of Luxembourg for providing the very necessary computing power and answering my many questions.

Finally, my friends and family both here and back in the USA have kept me going throughout. Thanks to Dan, Kristine, and John for letting me ramble and vent. Without your encouragement I would not have made it through the first year in Luxembourg. Most importantly though, I could never have done this if my mother and sister had not supported me and taken care of my horse! Thank you both very much.

List of Abbreviations

BAM/SAM	Sequence Alignment/Map format
BFB	Breakage-fusion-bridge
CIMP	CpG island methylator phenotype
CIN	Whole chromosome instability
CML	Chronic myelogenous leukemia
COSMIC	Catalogue of Somatic Mutations in Cancer
CpGI	CpG Islands
CSC	Cancer Stem Cells
DE	Differential Evolution (algorithm)
DSB	Double-stranded-break
EA	Evolutionary Algorithm
EC2	Amazon Elastic Compute Cloud
EM	Expectation Maximization (method)
GA	Genetic Algorithm
GVF	Genome Variant Format (files)
HPC	High performance computing
HTS	High-throughput sequencing
ICGC	International Cancer Genome Consortium
ISCN	International System for Human Cytogenetic Nomenclature
LD	Levenshtein Distance
LOH	Loss of heterozygosity
MSI	Microsatellite instability
NCD	Normalized Compression Distance
SNP	Single nucleotide polymorphism (also referred to as SNV)
SNV	Single nucleotide variation
SV	Structural variation
TCGA	The Cancer Genome Atlas

TABLE OF CONTENTS

FIGURES & TABLES

FIGURES

TABLES

SUMMARY

As a genomic disease cancer is unique in that the entire genome can be highly unstable, with new mutations accumulating at a rapid rate and massive alterations to the chromosomal structure. Structural aberrations can be highly significant to a patient's disease, resulting in aberrant proteins that can drive a cancer to progress faster or metastasize. Such aberrations may also have more subtle effects, enabling the cellular population to more rapidly develop drug resistance or simply generate highly diverse populations within a tumor making targeted therapies less effective.

In fact it is these diverse or heterogeneous cellular populations, with highly mutated and frequently structurally aberrant genomes, that make understanding the extent of a tumor genome's variation so challenging. Large scale sequencing efforts through the Cancer Genome Atlas and the International Cancer Genome Consortium have sequenced thousands of cancer genomes, and while small-scale variants have enabled researchers to begin to trace the evolutionary history and diversity of tumor genomes, large-scale structural variations have continued to be difficult to identify.

Current methods and technologies for short-read sequencing generally rely on fitting genomes to a single reference assembly that is assumed to be representative of all individuals. Tumor genomes, which consist of heterogeneous cellular populations with unique aberrations can vary significantly from a 'normal' genome. This means that such single references are poor representations of a cancerous cell population, and so methods that rely less directly on the reference offer better opportunities to investigate these aberrations.

In this project, a new method for large-scale structural variant identification, called *MultiSieve*, is proposed. This method uses prior knowledge to generate and test multiple references for each patient genome. Validation using simulated data establishes the utility of the method, and a comparison with commonly used methods demonstrates that *MultiSieve* is capable of finding variations often missed by traditional methods and that there are likely to be more structural variants in patients than have been identified previously.

CHAPTER 1 INTRODUCTION & LITERATURE REVIEW

Cancer is one of the leading causes of death worldwide today. In the United States and Europe it is second only to cardiovascular disease in the number of deaths each year. It is not a new or modern disease either. Evidence of metastatic disease in mummies and in writings from 1200 BC in Ancient Egypt, Neanderthal fossils with evidence of bone cancer, and the many other species that develop tumors show that this disease has always been with us. In the 18th century doctors made several important observations about possible causes of the disease linking it to tobacco use, hormonal changes, and environmental exposures to carcinogens in the workplace. However, for centuries the only available treatment was surgical removal. By the early 1900's radiation was being used to kill (and create) cancerous cells, and the 1950's saw the introduction of the first chemotherapies. Researchers have since developed many different therapies to treat the various cancers.

In recent years the aim of cancer research has focused on the understanding of early development and progression of the disease for the development of earlier treatments, avoiding the challenges associated with late stage or metastatic disease. As detection methods have improved, enabling doctors to recognize cancer in earlier stages, available therapies have also become more effective. It has also become increasingly clear how challenging identifying early stage cancer can be both in regards to physical location (e.g. deep tissues such as ovarian versus breast) and in identifying accurate biomarkers at early stages. Complicating treatment further is that tumors also display significant heterogeneity in cellular morphology, genomic stability, and drug responses as well as in their progression.

Despite this heterogeneity all cancers share a common mechanism of development. Each disease is the result of an evolutionary process of mutation (random or environmentally induced) and selection from within the tissues of the organism. Selection may be simply due to deleterious effects of the mutations on single cells, or driven by immune targeting of mutated cells. Eventually some set of mutations in a cell will enable it to proliferate unchecked generating a tumor, and ultimately invading other tissues to metastasize. This process continues within the tumor, continuously generating cellular populations with novel mutations. The result is heterogeneity at every scale from the tissues involved in the disease to the cellular makeup of the tumor itself. Determining whether, or to what extent, the large genetic profile of mutations contributes to tumor progression or drug response continues to be a central question in cancer biology.

Finding and identifying mutations that can explain a given phenotype is the primary goal in the search for genomic causes of diseases. Most often this search has focused on

mutations within single or specific genes and attempting to link these to a disease phenotype. In some diseases this is relatively simple, enabling single genes to be mapped to known diseases or disorders as in cystic fibrosis (e.g. CFTR gene) (Hamosh et al. 2005). Many of these genes and mutations have been identified through a process of sifting through the sequence of patients to identify rare mutations to a gene or functional alterations to a protein. A number of these types of mutations have been identified as increasing the risk of developing cancer (e.g. BRCA1, TP53) and the search for other such 'driver' mutations has led to the identification of hundreds of potential genetic drivers and many thousands of mutations (Pleasance, Cheetham, et al. 2010).

It has become increasingly clear in the process of searching for mutational drivers that cancer is a highly complex and diverse disease with many overlapping types and subtypes. There exists large differences even between patients sharing the same disease (Alexandrov et al. 2013). Signatures of different mutational processes have been found across cancer types suggesting that similar mechanisms may be driving the underlying complexities, at least with regards to single nucleotide mutations. Furthermore, these signatures display mutational similarity between cancers such as esophageal and colon cancers, or lung and bladder cancers, despite being phenotypically different cancer types (Lawrence et al. 2013). Complicating the picture is that small base pair alterations are not the only form of mutation found in the cancer genome. Epigenetic changes, whole chromosome duplications or deletions, and large-scale rearrangements of chromosomal regions resulting in massive differences from the normal human genome are common across cancer types. These are often (though not always) associated with poorer outcomes, found in later stage disease, or related to drug resistance.

The search for underlying genomic causes has been accomplished through various approaches ranging from karyotyping tumors to genome-wide association studies (GWAS) to identify candidate or driver genes. High-throughput sequencing (HTS) revolutionized cancer genomics research by providing the technological ability to rapidly generate high quality sequence data. Sequences from normal tissue and the matched tumor in a single patient have enabled researchers to search for mutations that are unique to the tumor and identify distinct mutational signatures (Alexandrov et al. 2013). Extremely high-depth sequences have also helped to unravel some of the evolutionary history of a tumor. However, there are still significant limitations to the identification of large-scale structural aberrations common to cancer genomes.

In the following sections these issues will be detailed. Section 1.1 (Genome Instability & Chromothripsis) describes the types of genomic instability in the cancer genome and what mechanisms may be responsible. Section 1.2 (Evolutionary Progression) will go on to discuss the development of tumor heterogeneity and the evolutionary processes

responsible. Section 1.3 (Genome Sequencing) will provide a detailed discussion of sequencing technology and the issues with structural variation identification. Finally, section 1.4 (High Performance Computing for Genome Analysis) will discuss the recent technological advances in computing that are being used to enable analysis of large sequencing datasets.

1.1 Genome Instability & Chromothripsis

Genome replication and division is generally an accurate process with a point-mutation rate of only 0.77 x 10^{-9} per site per cell division (Lynch 2010). Errors in chromosomal segregation are even more uncommon, occurring at a rate of about 1 in 100 cell divisions, as they are typically detrimental to a cell that is not already neoplastic (Thompson and Compton 2010; Manning, Benes, and Dyson 2013). In contrast, cancer genomes often have a high mutation rate (Loeb 2001), dependent on carcinogen exposure (e.g. tobacco, aflatoxins, radiation), the stage of the cancer, and the tissue type (Stratton, Campbell, and Futreal 2009; Berger et al. 2011). In neoplastic cells the mutation

Table 1 Common Structural Variations

Feature	Description
Whole genome duplication	Duplication of entire chromosomal set, often the result of cytokinesis failure
Chromosomal loss/gain	Gains or losses of individual chromosomes
Chromothripsis	Catastrophic rearrangements involving two or more chromosomes
Chromoplexy	Chained structural rearrangements that appear to be gained sequentially
Translocations	Karyotype level chromosomal rearrangements between two chromosomes up to whole arms
Copy number gain/loss	Duplication/deletion of genomic regions caused by structural rearrangement or replication errors

Common structural variation features of a cancer genome. These features can radically alter the genomic landscape, result in gene-fusions that alter protein expression, and influence gene expression that is dosage dependent.

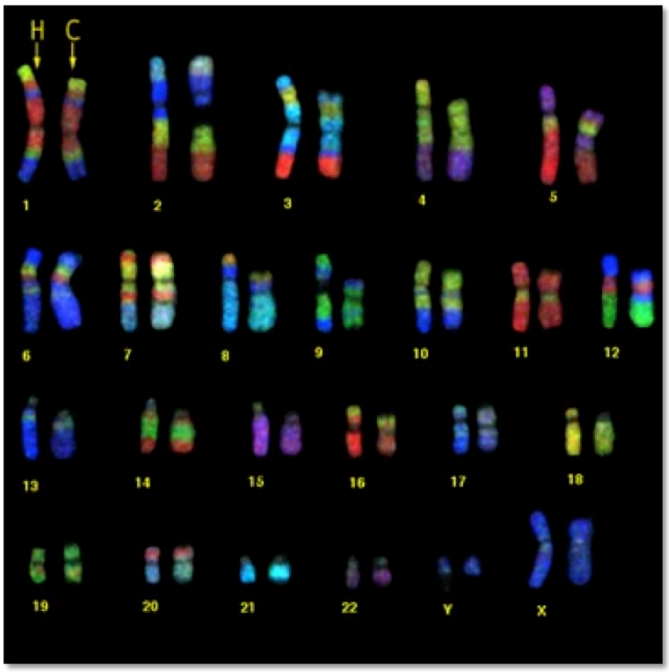
spectrum can be broad as well, including a variety of both small and large (see Table 1) scale changes from single nucleotide variations (SNV) and small indels, to translocations, inversions or deletions of large chromosomal segments. In addition to mutational changes, tumor genomes often exhibit whole-chromosome and whole-genome aneuploidy. These two types of instability contribute to tumorgenesis, and provide genetic diversity for drug resistance, though they also potentially compromise the viability of a cellular population (Janssen and Medema 2012). Understanding how these two mechanisms contribute to the process of tumorgenesis may be key to halting or reversing

disease progression.

1.1.1 Numerical Instability

The most common form of genomic instability in tumors is whole-chromosome instability (CIN), also known as aneuploidy, where cells may contain non paired (for most animals) chromosomes. At the most basic level this occurs where chromosomes fail to segregate correctly during mitosis due to microtubule defects, mutations resulting in dysregulation of mitotic checkpoint genes, or cell fusion. The result can vary in a single geome with an incorrect number of chromosomes, both too few or too many, to the entire complement of chromosomes being fully duplicated.

Figure 1 Human & Chimpanzee Karyotypes



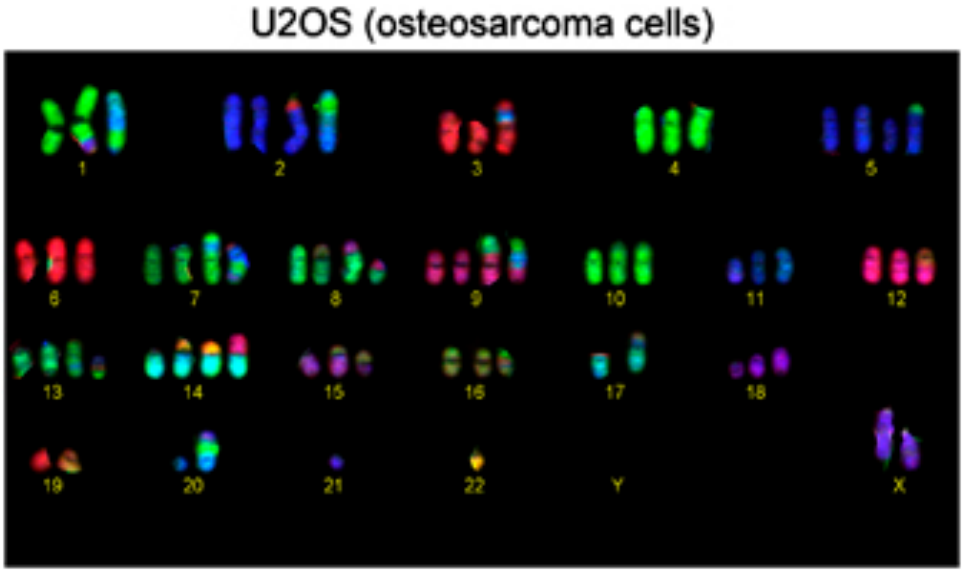
Human (H) and chimpanzee (C) karyotype side-by-side shows the primary difference in human chromosome 2 where two chromosomes from the great ape lineage fused. Image from http://www.nationalmediamuseum.org.uk/.

Evolutionarily aneuploidy may have driven speciation as seen by closely related species containing different numbers of chromosomes. For instance, chimpanzees and other great apes have a diploid complement of 48 chromosomes while humans have 46. This difference is due to the formation of chromosome 2 in humans from the fusion of two smaller chromosomes 2A and 2B (see Figure 1) in the great apes (IJdo et al. 1991). Other closely related species display similar ploidy changes: domestic horses have 64 to the Przewalski's horse 66 or the donkey's 62; domestic dogs and wolves have 78 while their close relatives the maned wolves have 76; or African elephants 56 chromosomes to the wooly mammoth's 58. In embryonic development of most animals however, aneuploidy is often lethal or significantly deleterious to the organism (including in the crossbreeding of close species with different numbers of chromosomes). Evidence from human Trisomy 21 points to altered transcriptional regulation due to gene dosage effects (FitzPatrick et al. 2002), though these effects appear to be subtle, as a possible explanation.

In cancer aneuploidy is common and possibly useful to tumor cell populations.

Aneuploidy in cancer has been correlated with drug resistance and advanced tumor grades, and has been found in most cancer types (R. A. Burrell et al. 2013). It has been difficult to identify specific causes in cancers due to the frequent co-occurrence of numerical CIN and structural instability resulting in massively altered chromosome structures. One important note however, is that while heritable mutations in the DNA repair pathway genes involved in chromosome segregation have been associated with familial cancers (Kim et al. 2012), in most spontaneous tumors these mutations appear to be mutually exclusive with the appearance of aneuploidy. This has been shown in the few solid tumors that exhibit no aneuploidy as they display a much higher rate of sequence level mutation (Cheng et al. 2008).

Figure 2 Osteosarcoma Cell Line Karyotype



Karyotype of osteosarcoma cell line displaying both aneuploidy in all chromosomes as well as structural rearrangements. Chromosomes 9 and 14 are an especially clear example of translocation. Image from (Janssen and Medema 2012).

Since mutation in the pathways and genes that control chromosome segregation (e.g. mitotic checkpoint, microtubule-kinetochore formation, and chromatid cohesion) does not appear to be directly involved in cancer related CIN, a new concept was introduced: oncogene-induced mitotic stress (Duijf and Benezra 2013). Here altered expression of tumor-suppressor genes or common oncogenes may affect chromosomal segregation either directly or indirectly. For example, in retinoblastoma the combined loss of functional tumor-suppressors in the p53 and pRB pathways showed a strong correlation with higher rates of CIN (Manning, Benes, and Dyson 2013), while the loss of just p53 was not sufficient to increase the rate of CIN. This could help explain why most cancers will eventually develop aneuploidy as well as the order in which it occurs. As loss of either p53 or pRB is widespread across cancer types, acquisition of a mutation inactivating one

where the other already has defects may tip the balance of segregation errors.

How CIN affects drug resistance and tumorgenesis is not entirely clear. One proposed mechanism is induced instability at both the sequence and structural levels due to ongoing mitotic defects, referred to as the 'mutator phenotype' (Loeb 2001). One of the primary effects of CIN may be in altered gene dosage or in the loss of heterozygosity. Although, since the gain or loss of entire chromosomes means many genes are affected this seems likely to be detrimental to cellular populations, especially as cancer cells tend to show multiple chromosomal gains and losses.

1.1.2 Structural Instability

In addition to numerical instability large-scale structural instability is found across many cancer types. These range from intra-chromosomal inversions or duplications of several thousand base pairs, to translocations involving two or more chromosomes resulting in chromosomes that are visibly altered (see Figure 2). These rearrangements may result in alterations to gene expression due to interruption of transcriptional regions (Shigesada, van de Sluis, and Liu 2004), or gene fusions that produce an altered protein product (D. R. Robinson et al. 2011).

Table 2 Known Gene Fusions

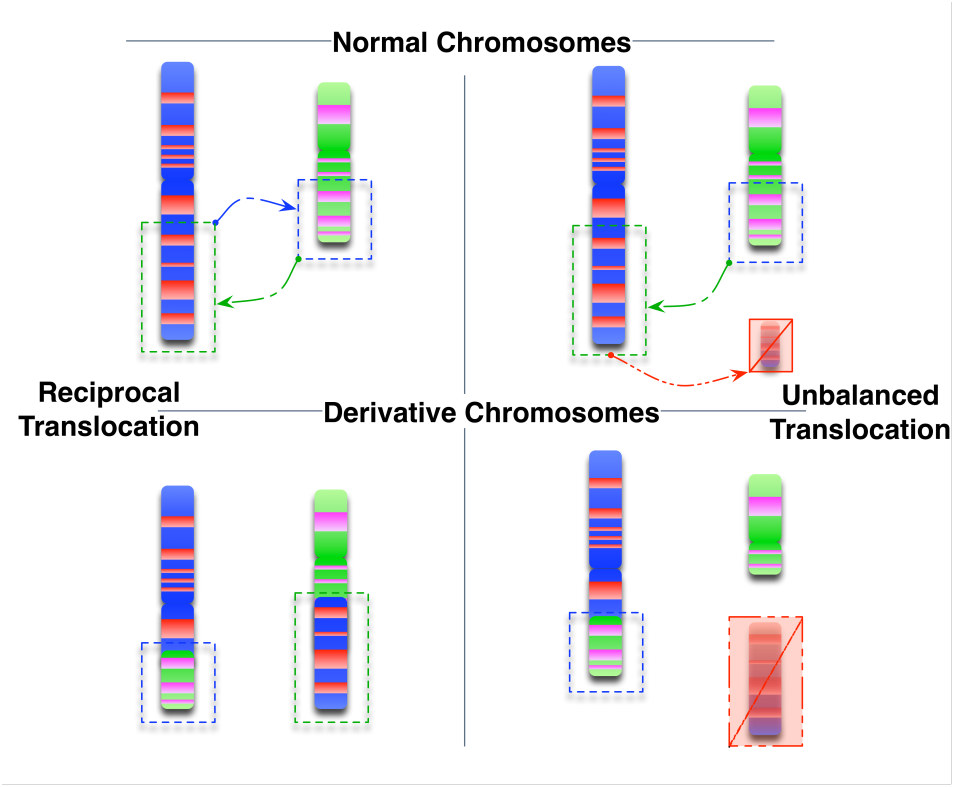
Fusion Genes		Cancer	Regions	
ETV6	NTRK3	Breast cancer	12p13	15q25
TFE3	PRCC	Renal-cell carcinoma	Xp11	1q21
TET1	TP53	Kidney clear cell sarcoma	10q22	17p13
PAX3	FKHR	Alveolar rhabdomyosarcoma	2q35	13q14
EWS	CHN	Myxoid chondrosarcoma	22q12	9q22-31
EWSR1	FEV	Ewing's Sarcoma	22q12	2q36
DEK	NUP214	Acute lymphoblastic leukemia	6p22	9q34

A few examples of known gene fusions across cancer types, there are many more listed by (Nambiar, Kari, and Raghavan 2008).

In 1961 Nowell and Hungerford described the first translocation in multiple chronic myelogenous leukemia (CML) patients between chromosomes 9 and 22 (Nowell and Hungerford 1961), later designated the Philadelphia chromosome. While this was the first structural aberration described in cancers, in 1914 Theodore Boveri had hypothesized that such abnormalities could be the cause of tumor development long before cytogenetic techniques enabled researchers to visualize them. In the years since Nowell's description of the Philadelphia translocation, structural changes visible at the level of the karyotype have been described in all cancer types.

The result of the Philadelphia translocation is a fusion point between the breakpoint cluster region (BCR) on chromosome 22 and the Abl oncogene on chromosome 9, creating the Bcr-Abl gene fusion that encodes for a mutant tyrosine-kinase protein that both increases the rate of cell division and inhibits DNA repair, inducing genomic instability (Look 1997). This aberration is not specific to CML either. It has been found in acute lymphoblastic leukemia (ALL) as well as acute myelogenous leukemia making it a hallmark of hematopoietic disease. The Bcr-Abl fusion protein also provided a therapeutic target for the development of tyrosine-inhibitor therapies (Kaelin 2004).

Figure 3 Translocations



Translocations can either be reciprocal (left side) resulting in two derivative chromosomes with the same total genetic material, or unbalanced (right side) resulting in the loss of some portion of one or both chromosomes. Unbalanced translocations can involved 2 or more chromosomes in more complex rearrangements with frequent losses of some sequences that do not recombine elsewhere in the genome.

Since then gene-fusions resulting from translocations have been documented across various cancer types (see Table 2), however translocations that do not result in known mutated protein products are far more common numbering in the hundreds to thousands of translocations depending on the cancer (Mitelman, Johansson, and Mertens 2007). Such translocations may be balanced or reciprocal where two chromosomes swap some genetic material creating two fusions, but no net sequence loss or gain. These reciprocal translocations are not uncommon in the general population as they can occur in the

development of gametes and may not cause deleterious effects in fetal development. In cancer however, translocations are both common and complex with unbalanced rearrangements that significantly alter chromosomal structure (see Figure 3).

Complex structural rearrangements involving tens or hundreds of chromosomal regions may be gained in a one time event termed *'chromothripsis'* (Stephens et al. 2011), from multiple sequential events occurring in the same regions known as *'chromoplexy'* (Baca et al. 2013), or in a gradual accumulation over time. The result of any of these events is massive alterations to the genome itself and has been documented in most cancers. These are thought to promote tumorgenesis and drug resistance by providing genetic diversity to the tumor with adaptations that are specifically important in cancer such as loss of tumor suppressor genes (Baca et al. 2013), acquisition of drug resistance (R. a. Burrell et al. 2013), apoptosis resistance (Stephens et al. 2011), promotion of angiogenesis (McBride et al. 2012), or the ability to evade the immune system (Janssen and Medema 2012). In fact, the complexity and number of large-scale events increases with tumor grade in most cancers (Mitelman, Johansson, and Mertens 2007; Duijf and Benezra 2013; Baca et al. 2013) supporting the assumption that these structural events help to support the tumor or even lead to metastasis.

Gradual accumulation of somatic mutations over an individual's lifetime has been the commonly accepted model for the development of a tumor genome (Stratton, Campbell, and Futreal 2009). This model argues that as the cell acquires small mutations over time due to damage that was not repaired and therefore becomes fixed in the genome, it eventually tips the balance into unrestrained growth. The rate of mutation acquisition may differ over time or in different tissues. Mutagenic exposures such as tobacco or UV radiation can increase the rate of mutation, while rates may be lower in the slower growing tissues of colonic crypts. It is clear that point mutations can accumulate in cells over the lifetime of an individual without resulting in cancer (Holstege et al. 2014), and so identifying which mutations in a given tissue may be the driver towards cancer continues to be difficult. However, lifetime acquisition of mutations only applies to small mutations affecting a few nucleotides at a time, large-scale rearrangements of the genome are unique to tumor progression.

Sudden changes in the genome of tumors have been documented across various cancers types in patients (Zack et al. 2013) and is echoed throughout the cancer cell lines. Unless such changes are immediately lethal to the cell, these rearrangements are also passed to daughter cells, creating new clonal populations containing structurally altered genomes. These *'chromothriptic'* events appear to occur suddenly in a catastrophic event and result in multiple rearrangements, rather than progressive accumulations (Stephens et al. 2011; Przybytkowski et al. 2014). Such catastrophic

events result in fewer changes to copy numbers, as chromosome fragments are either lost resulting in lower copy numbers, or retained in a new derivative chromosome resulting in little or no change to the copy number. This can make such events, and their derivative sequences, difficult to identify.

Another genome remodeling process appears to be at work in some tumor genomes that results in a complex arrangement of balanced translocations. First described in prostate cancer (Berger et al. 2011), a complex and potentially progressive pattern of breakage and recombination was shown in tumor samples. This pattern, termed *'chromoplexy'*, resulted in no loss of sequence material as the derivative chromosomes were all balanced translocations. Interestingly the prostate genomes showed a common pattern where the breakage and recombination occurred near the same regions across multiple patients. These suggest that the rearrangements were part of a chain of events, where each event caused a dysregulation that drove the next event in the chain. Evidence for this chained event was shown through analysis of the sub-clonal alterations within the tumor (Baca et al. 2013) and was related to higher grade, clinically aggressive tumors.

While the mechanisms for the development of patterns of chromothripsis/chromoplexy is still unclear, general structural instability is thought to arise primarily through incorrect non-homologous end-joining while attempted repairing of double-stranded breaks (DSB) (Moynahan and Jasin 2010). A number of environmental causes of DSBs have been shown to result in structural aberrations including UV radiation, chemical mutagens, and ionizing radiation (potentially the cause of chromothripsis in non-hematopoietic cancers). DSBs may also be more common in genomes with microsatellite instability (MSI) due to higher sensitivity in the DNA repair pathway to MSI mediated mutation (Bilbao et al. 2010; H.-R. Li 2004). It has also been shown that translocations acquired in DNA repair pathway genes may enable the accumulation of additional translocations in a series of catastrophic alterations due to deregulation of the repair process. Other processes may also drive the accumulation of aberrations including the breakage-fusion-bridge (BFB) cycle. BFB can cause regions with DSBs to fuse, resulting in aberrations such as dicentric chromosomes (e.g. a derivative chromosome that includes the centromeres from two chromosomes), driving further breakage events through improper microtubule attachments during mitosis (Guerrero et al. 2010).

It also appears that some regions of the genome are more prone to DSBs due to known fragile loci in the chromosomes. This has been seen in hematopoietic cancers which have a high rate of structural aberration that is likely due to the rapid proliferation of lymphocytes combined with pre-programmed genome remodeling that is part of the specific cell lineage, making these cells especially sensitive to errors in the DNA break/repair process (Barlow et al. 2013). Small-scale structural aberrations, such as

localized deletions, at known fragile sites on the genome (Ried 2000), or clusters of large homozygous deletions may also contribute to large-scale remodeling or chromothriptic events (Bignell et al. 2010).

That these structural rearrangement events may happen at various times in a tumor cell population's lifetime suggests that structural instability can generate as many driver mutations in the development of a malignancy (e.g. CML and Bcr-Abl) as neutral or passenger mutations throughout the replicative lifetime of the clone (Pleasance, Stephens, et al. 2010). It appears highly likely that even structural variants that occur close to gene boundaries are neutral in a given sub-clonal population as the proportion of breakage/recombination events is significantly higher than is the identification of recurrent fusion genes across patients even within the same tumor type (McBride et al. 2012).

There are a number of underlying causes of structural aberrations in tumors, and these aberrations occur in all cancer types. It is for this reason that better methods are needed to identify, characterize, and associate them with disease prognosis. These large-scale structural variations (SV) are often associated with the progression of malignancy and even metastasis in cancer. The next section discusses these underlying driving forces in more detail.

1.2 Evolutionary Progression

Complex multicellular life is possible only through the cooperation and tight control of the billions of cells within an organism. Controls are needed that enable differentiation, tissue specialization, intra-cellular communication, and some ability to recognize other cells as 'self'. When these controls fail a single cell can grow and divide unchecked, resulting in a tumor. From that first transformed cell to metastatic disease cancer develops through an evolutionary process of mutation, expansion, and population selection.

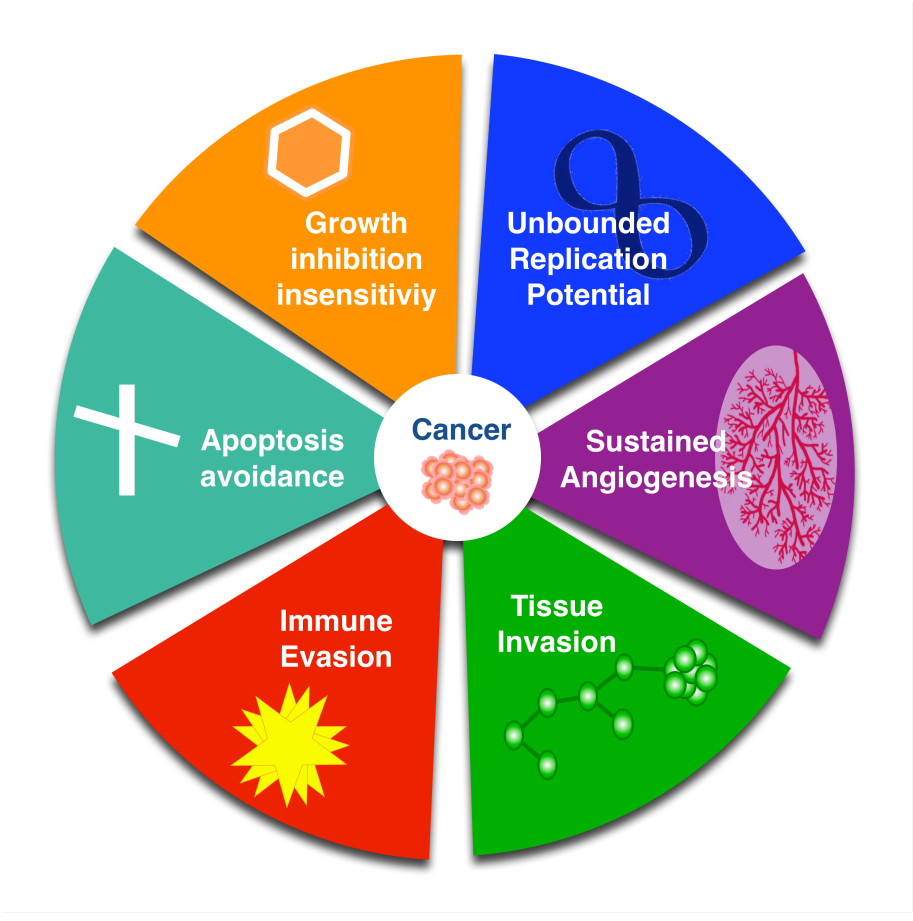
Cancer evolution takes places within a tissue ecosystem, which is tightly regulated to optimize the function of the tissue itself. During the early evolution of a tumor that environment may limit (e.g. colon cancer develops slowly due to the slow process of crypt cell proliferation) or enable expansion (e.g. small-cell lung cancer is typically fast growing due to carcinogen exposure). Specific mutations that provide competitive advantage to a specific cell or population may also enable a more aggressive progression (Stratton, Campbell, and Futreal 2009). For instance, mutations in the RAS pathway, specifically the KRAS oncogene, have been identified across a variety of cancers and the order in which the mutations occur is predictive of cancer progression: in colorectal cancer KRAS2 mutations are unlikely to progress to malignant disease unless APC was inactivated first (Vogelstein and Kinzler 2004; Attolini et al. 2010); in lung adenocarcinoma tumorigenicity requires both KRAS and p53 mutations (Meacham and Morrison 2013). These mutations

provide selective advantage to the clones and are considered the 'drivers' of tumorgenesis.

1.2.1 Hallmarks of Cancer

Whether the mutations that are initially acquired by a cell are due to errors in replication and repair, environmental mutagens, or heritable mutations to tumor suppressor genes, only begins to matter when those mutations alter the function of specific pathways. The progression from normal healthy cells to cancer is not spontaneous, occurring from one cell division to the next. Instead each mutation acquired may provide the cell with advantages compared to the cells within the local environment. In order for a single cell to generate a population of cells that will become a tumor the acquired mutations must alter the function of several pathways that are critical to tissue maintenance in an organism (see Figure 4):

Figure 4 The Hallmarks of Cancer



Tumorgenesis is a multi-step process that involves alterations to these six general pathways: growth inhibition, apoptosis, replication potential, angiogenesis, tissue invasion, and immune system evasion (and sometimes immune recruitment).

- *Growth inhibition insensitivity.* In normal tissues cellular proliferation is a rigidly controlled mechanism. Various signals maintain tissue homeostasis by keeping cells in a quiescent state or preventing proliferation in cells that have entered the growth phase of the cell cycle. One of the earliest known tumor suppressor genes, pRB, inhibits transcriptional activities of the E2F transcription factor leading to cell cycle arrest (Bracken et al. 2003). Dysregulation of this pathway prevents inhibition of the E2F family of genes, allowing cells to progress from G1 to S phase. Other paths to dysregulating the pRB pathway are likely to be involved as well, such as preventing the expression of cell adhesion molecules that are involved in growth suppression.

- *Apoptosis avoidance.* Both a high cellular proliferation rate and resistance to apoptosis signals enables the development of tumors. Disruption of the p53 pathway which induces apoptosis in the presence of DNA damage enables tumorgenesis and is involved in a high proportion of cancers (Hickman 2002). This is also one of the drivers of genomic diversity in tumors as cells with DNA damage continue proliferating. However, resistance to apoptotic signals does not necessarily result in immortal cells or cellular populations. One example is necroptosis (Vanden Berghe et al. 2014), a regulated form of necrosis (cellular injury), which is mediated by inflammatory pathways (TNF signaling) rather than cell cycle pathways.

- *Unbounded replication potential.* Normal cells have a limited number of replications - in cell culture 60-70 doublings - once that limit is reached the cellular population stops growing. While cellular populations with inactivated tumor suppressors (e.g. pRB and p53) may continue past this limit, the population eventually reaches a so-called crisis state resulting in cell death throughout the population and the appearance of a few cells that have acquired the ability to replicate without limits (Hayflick 2000). As most tumor cells in culture appear to be immortal this may happen early in tumor progression, suggesting that the limit applies to the tumor precursor cells.

- *Angiogenesis.* Basic cellular function is supported by the oxygen and nutrients supplied to all cells in the body by the vascular system. Organs within the body develop with their own supporting blood vessels to deliver these nutrients directly. As tumorgenesis progresses the abnormal cells must also access these critical nutrients. This requires that the cells within the tumor be capable of promoting angiogenesis, through disruption of angiogenesis pathways such as VEGF (vascular endothelial growth factor), TGF (transforming growth factor), and TNF

(tumor necrosis factor). While angiogenesis is necessary to the organism for tissue repair or reproduction, uncontrolled angiogenesis is required for tumor development (Nishida et al. 2006). Disruption of the regulatory pathways of angiogenesis supports tumor progression in solid tumors, while inhibition of these pathways appears to slow tumor growth.

- *Immune system evasion.* The immune system plays a key role in the identification and elimination of cells that have transformed. This anti-tumor immune response controls early development of a tumor and continues to attack tumor cells using tumor-specific T cells (CD4+, CD8+), though this is complicated as most cancer cells lack MHC-II and so the response primarily relies on antigen presenting cells (Corthay et al. 2005). Though the mechanism is currently unclear it is apparent that in order for proliferation to continue tumor cells must have evaded or resisted the immune response. Many tumors continue to elicit an immune response through the presence to T cells within the tumor microenvironment. As there is such continual anti-tumor activity, tumor cells must also continually to avoid detection in order to proliferate and eventually invade other tissues (Gajewski, Schreiber, and Fu 2013).
- *Tissue invasiveness.* In order for a tumor to spawn secondary tumors, or metastasize, the capability for tumor cells to invade adjacent tissues and travel to new sites within the organism is necessary. That this capability tends to be acquired late in tumor stages suggests that, in part, the tumor may occupy an environment that is too competitive to support the continual development of new cellular populations or that it is simply too large for the poorly coordinated vasculature to provide for. It is not entirely clear what specific pathways are altered to allow a cell to acquire invasive/metastatic capability. Those involved in cell adherence have been identified in metastatic epithelial cancers, as well as alterations to expression of integrins, and extracellular proteases (Hanahan and Weinberg 2011).

The specific mechanisms and acquisition timing of each of these capabilities may differ between cancer types, patients, or even from one tumor to another. Additionally, it is not necessary that each cellular population within a tumor have acquired all of these capabilities. It appears, for instance, that only some cells within a tumor have invasive or metastatic potential while other populations that may have arisen entirely within the tumor may not have the ability to evade anti-tumor T cells. However, each of these capabilities is ultimately critical to enable tumorgenesis and eventual metastasis.

As the path to gaining the 'hallmarks of cancer' involves significant disruption to the genome, it ultimately gives rise to genomic instability. This instability most often results in

the development of many different tumor subpopulations, an evolutionary process of clonal expansion through relative fitness advantage provided by mutations. Each of these subpopulations carry both driver mutations (e.g. mutations to cellular process important to tumorgenesis) and neutral mutations that may have no effect in a given population at that time, but are carried from within the population during successive clonal expansions. Individual tumors may contain many thousands of mutations relative to the normal genome, and most are presumed to be neutral. As many hundreds of genes are found to be mutated in cancers with no clear causative influence (e.g. olfactory receptor genes), have high mutation rates across cancer types, and have rates that appear dependent on the length of gene or timing of replication (Lawrence et al. 2013) this indicates the acquisition of a large number of neutral mutations alongside adaptive mutations.

Genomic instability and the varying mutation rates across subpopulations results in dynamic and heterogeneous cellular populations within the tumor. Once the threshold of malignancy has been crossed the fitness advantage for tumor cells is not relative to normal tissues, but relative to subpopulations within the tumor. The relative abundance and mutational profiles of the populations varies both by cancer type as well as disease stage, though it appears that most cancers have a numerically dominant subpopulation (Nik-Zainal et al. 2012; Jiao et al. 2014). This dominant population is not predictive of the populations that will be important to the development of metastasis (Hou et al. 2012), instead it is often simply the proliferative population responsible for the bulk and expansion of the tumor.

Understanding this evolutionary progression has become increasingly important in the ongoing efforts to develop more accurate diagnostics and targeted therapeutics. Tumor development appears to be highly dynamic, and can arise due to monoclonal or polyclonal somatic mutations (Stephens et al. 2011; Visvader 2011). This leads to one of the primary difficulties in understanding the mutations that drive cancer development and progression: tumor evolution is a dynamic process that varies from one tumor to another, even within a single patient.

1.2.2 Tumor Heterogeneity

It has been well established at this point that cancer is not a single disease, that each tumor type has both distinct morphology and progression, and that even the same type can vary significantly between two patients. Complicating this further is that tumors themselves are often highly heterogeneous with multiple populations of proliferating, quiescent, tumorigenic, and non-tumorigenic cells. Understanding this heterogeneity within individuals is likely to provide the key to targeted or personalized therapies.

At the point where a tumor is clinically recognized the ‘cell of origin’ (Visvader 2011),

or the first normal cell to acquire the minimum necessary hallmarks of cancer has been long eclipsed by the proliferating populations around it. However, depending on how the subsequent populations were generated it may still be possible to identify the mutations that enabled acquisition of these capabilities. Identifying which drivers are important to each cancer type can help to determine disease progression. For example, mutations in the RAS pathway (specifically the KRAS oncogene) have been identified as significant in colorectal, lung, and pancreatic cancers. Simply finding the mutations is not enough to be predictive of tumor progression, instead the tissue and order in which the mutations occur must be known. In pancreatic cancer, primary mutations to KRAS2 appears to drive the development of disease, while in colorectal cancer KRAS2 mutations are unlikely to progress to a malignant disease unless APC was inactivated first (Vogelstein and Kinzler 2004; Attolini et al. 2010). Finding the order of mutations would provide further understanding of the development of tumor populations.

One of the hurdles to the development of effective cancer therapies has been in determining which cells within a tumor need to be targeted. Therapies that attack the large proliferative population may slow down or even reverse the growth of the tumor, but leave in place a small population of cells that have acquired greater resistance to drugs or invasive potential that may result in new tumors. However, the question is greater than simply asking which population should be targeted. Instead this question relies on knowing how that heterogeneity may have developed in the initial tumor. Two different models have been proposed which suggest significantly different outcomes: the 'cancer stem cell' (CSC) versus 'clonal evolution' (Shackleton et al. 2009).

CANCER STEM CELL

This model starts with the assumption that all tumors are hierarchically organized with populations of tumorigenic cells that may have already acquired the necessary hallmarks (discussed above). These populations consist of cells that drive the development and progression of the tumor and eventual metastasis as well as non-tumorigenic cells. Under this assumption the tumorigenic cells are the source for all additional subpopulations within the tumor and when transplanted can cause disease individually (Ding et al. 2010; Stewart et al. 2011), while the non-tumorigenic populations have no capability to develop disease when transplanted.

It is important to recognize that the 'cancer stem cell' did not necessarily start as a normal stem cell, instead it proposes that once a cell has acquired the minimum hallmarks (e.g. apoptosis resistance, growth inhibition insensitivity, unbounded replication potential) it becomes the 'stem cell' that can differentiate into phenotypically diverse subpopulations, which are responsible for sustaining tumor growth. The hierarchical organization resulting

from differentiation would display genomically heterogeneous populations with clear precursors, tumors that reflect the hierarchy similar to organ development, and could be therapeutically targeted by focusing only on the subpopulations that are tumorigenic. Certain cancers have been shown to be consistent with this model, including some leukemia's (Fearon et al. 1986; Bonnet and Dick 1997), breast cancers (Al-Hajj et al. 2003), glioblastoma (Singh et al. 2004), colon (O'Brien et al. 2007) and ovarian cancer (Stewart et al. 2011). In each of these, transplantation assays showed that only a small subpopulation of cells within the tumor had tumorigenic potential and were therefore capable of generating new tumors, supporting the view that these cancers were hierarchically organized.

There is one major caveat to both this model and the supporting assays: that tumorigenic potential may be far greater than presumed due to the environment. It seems likely that tumor cell differentiation can provide some rudimentary organ supportive capabilities. This would suggest that unless tissue invasive capabilities have already been acquired (Ding et al. 2010), such cells are tumorigenic only in a specific environment. Transplantation assays would not provide that environment and as they require that human tumor cells be transplanted into mice the xenogeneic immune response (even immunocompromised mice will still have some response) will result in the destruction of some potentially tumorigenic cells. This further suggests that the specific microenvironment has significant influence on the fate of subpopulations and this has been shown in leukemia where only specific conditions result in the development of immortal cells and disease (Wei et al. 2008).

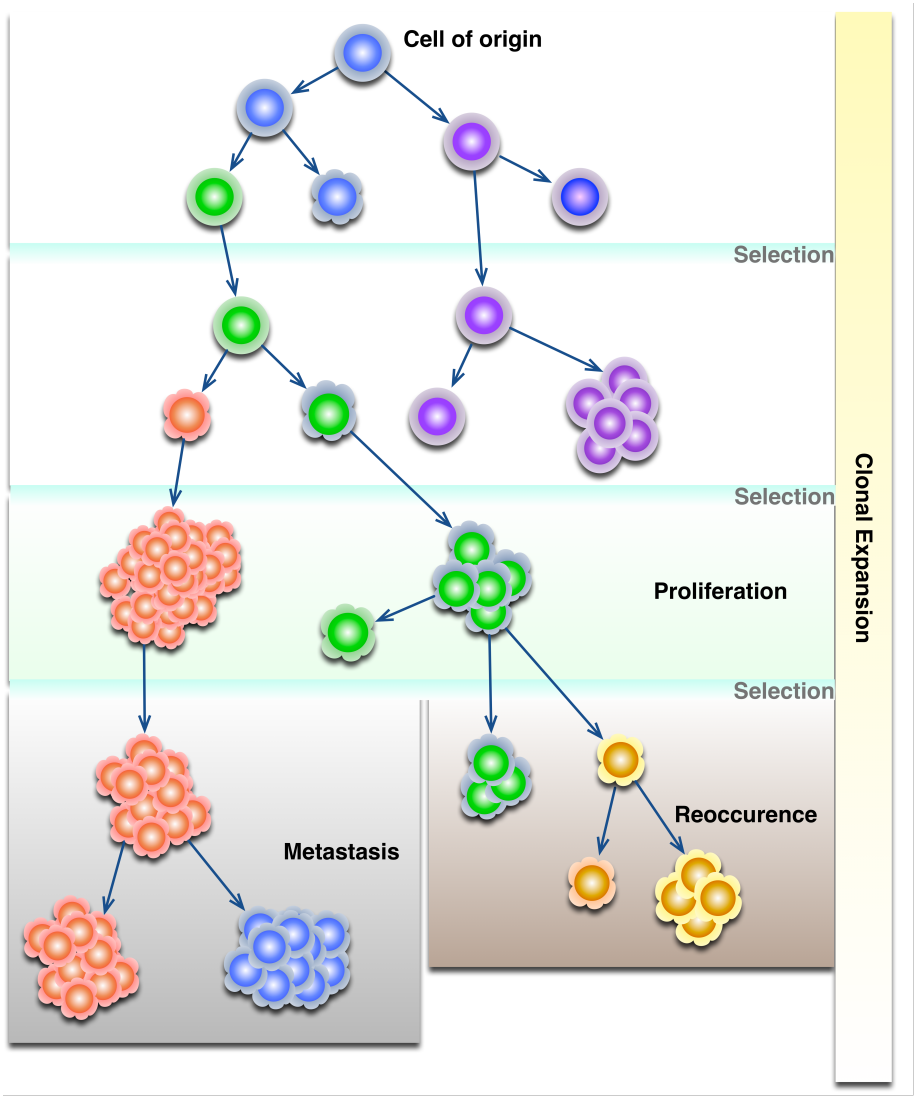
The most important consequence of the CSC model has been the assumption that the differentiation should result in markers (morphological or epigenetic) that could be used to separate the tumorigenic cells from the rest of the population. Unfortunately this has not been shown conclusively in part due to heterogeneity between patients making reliable marker identification difficult (Singh et al. 2004), as well as plasticity in some cancer cells that enable them to reversibly transition between states (Meacham and Morrison 2013).

Clonal Evolution

Clonal evolution was first recognized in cancers by Nowell (Nowell 1976) as a mechanism for tumor heterogeneity. This model recapitulates evolutionary pressures as each new clone acquires genetic or epigenetic changes (so-called 'driver' mutations) that enable expansion at different times under selective pressure (see Figure 5). In contrast to the hierarchical organization of the CSC model, clonal expansion will develop a branching pattern that may be locally hierarchical with different clones uniquely driving metastasis, relapse, or drug resistance. These branches represent dominance of individual clonal

populations at a given time, and points where selective pressures narrow the number of clones (Greaves and Maley 2012).

Figure 5 Clonal Evolution in Cancer



Cancer is an evolutionary process of mutation, selection, and expansion of clonal populations. Each color indicates clonally unique mutations.

Genetic heterogeneity between the clones is therefore a common feature in cancers and the primary reservoir for diversity within the tumor. While genomic evidence for clonal patterns in tumors is difficult to generate, as most analyses occur as a single snapshot in time of the tumor, both histopathology and high-throughput sequencing have provided a view of the process. Through tracking SNVs and other small mutations in sequence data it is possible to show clonal expansion and branching occurring in individual patients with breast cancer (Nik-Zainal et al. 2012), prostate cancer (Baca et al. 2013), myeloma (Bolli et al. 2014), and leukemia (Jiao et al. 2014). In each of these a dominant sub-clone is present representing the primary proliferating population at the time of analysis. Further

evidence has shown significant genomic differences between the dominant clone of the primary tumor and that of the metastasis, suggesting that only a minority of the population needs to develop metastatic capability.

The evolutionary progression of cancer means that within any given tumor biopsy there exists a mixed population of cells. Only a subset of this population will contain the necessary hallmarks of cancer, and are the possible targets of any clinical intervention strategy. Therefore it is recognized in this thesis that it is important that any genomic-based analysis should be able to identify these variants from complex mixed populations.

1.3 Genome Sequencing

The advent of high-throughput sequencing technologies (HTS) enabled a dramatic increase in available information from genomes. As of 2015 more than 170 eukaryotic species have been sequenced (http://www.ebi.ac.uk/genomes), as well as thousands of bacterial and viral species. The explosion in individual genomes has been even greater.

Less than decade after the completion of the first human genome sequence (International Human Genome Sequencing Consortium 2004) a project that aimed to sequence a thousand individuals across several populations has been completed (The 1000 Genomes Project Consortium 2010). At the same time the Cancer Genome Atlas (TCGA) Project was underway. TCGA aimed to sequence the germline (or 'normal') and tumor from 10,000 individuals across 20 different cancer types (The Cancer Genome Atlas 2008). The result would be 20,000 genome sequences for cancer analysis. Many other projects across human diseases, livestock, plants, and viruses have resulted in an explosion in the scale of genomic data available to the community. This has been driven by the advent of ever more efficient, rapid, and cheaper sequencing technologies. At the same time the continual improvement in sequencing technologies has meant that the raw size of genomic information for each individual sequenced has dramatically increased.

All of this has allowed many new investigations in the context of biomedicine, including: understanding parasitic diseases such as malaria (Gardner et al. 2002); identifying evolutionary and epidemiological dynamics of influenza that could be responsible for virulence (Rambaut et al. 2008); characterization of the complex microbial systems in the human body (Gill et al. 2006); and finding the variations that drive complex genomic diseases such as cancer (Campbell et al. 2008; Ley et al. 2008).

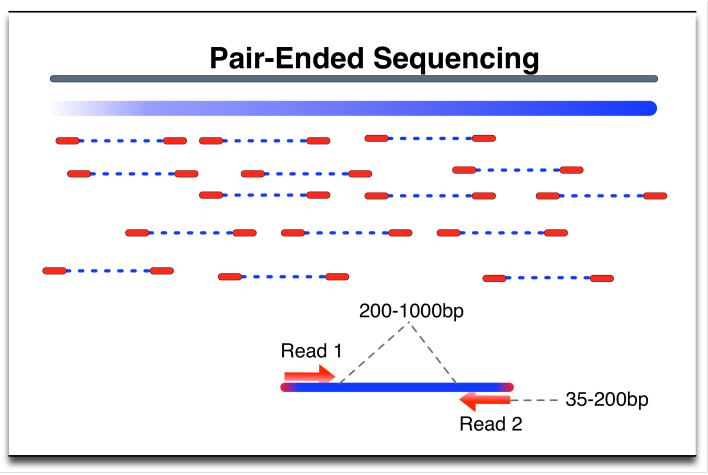
However, sequencing is not simply reading the output of a machine. The computational steps that are necessary in order to obtain information about individual variation and potentially deleterious mutations are both complex and incomplete. This is especially true in the context of complex rearrangements found in the cancer genome.

1.3.1 TECHNOLOGIES

Since the first HTS sequencing platform developed by 454 Life Sciences in 2006 was made commercially available many new instruments have been developed. Each iteration has seen the sequence output increase as the time and cost per individual genome has decreased. In fact the costs and time to sequence have dropped so dramatically - from 50,000USD and several weeks in 2010 to 1,000USD and a single day in 2012 – that many thousands of genomes have been sequenced. This means that the largest difficulties are no longer in obtaining a sequence, but in interpreting it.

The difficulties for interpretation are directly related to the process of sequencing. Currently sequencing platforms generate millions of short sequences, or 'reads', for each genome. These reads typically range from 35-200bp and, in paired-end sequences, with a gap or 'insert size' of 200-1000bp between each read (see Figure 6). These reads are generated at varying levels of coverage across the entire genome, allowing the reads to overlap so that where there is a gap in one pair of reads several other pairs will have covered it. A human genome sequenced at 30X coverage (100bp reads) will generate about 900 million reads, or 450 million pairs (Lander and Waterman 1988).

Figure 6 Pair-Ended Sequencing



Sequence reads are generated from both ends of the same fragment. For short-read sequencing the read lengths and gap (or insert size) between the reads will vary between platforms.

It is necessary to understand the limitations and inherent biases of the platforms when analyzing the resulting sequences. While experimental biases can be introduced at any step in the sample preparation process, genome amplificaiton is the most easily quantified and consistent source of bias. PCR amplification has been shown to bias sequence coverage in regions that are particularly GC rich/poor (Aird et al. 2011) at a higher rate than multiple displacement amplification methods (Pinard et al. 2006). These biases are of greater concern in smaller, low-complexity genomes such as plasmids. In human sequences the coverage bias appears to be more influenced by the length of the fragments. This in fact relates directly to the second source of bias in sequencing, which is that coverage across

the genome is directly related to the GC content so that GC rich regions show a decrease in coverage. However, this effect is also influenced by the library used (e.g. different libraries for the same sample will display different GC coverage) as well as by the specific lanes on the instrument (Benjamini and Speed 2012).

The last major limitation in short-read sequencing is related to the mappability of the reads themselves. Genomes are not random complements of base pairs with exceptions for identified functional regions. Instead due to duplications throughout the evolutionary history of a given genome, viral sequence inserts, and structural elements including centromeres/telomeres, there are large segments of any given genome that are repetative. In the human genome about 50% is estimated to be repetative, including portions made up of both long (centromeres) and short (ALU) repeat regions. This means that reads generated in these regions are likely to map to multiple locations simply due to their length (Treangen and Salzberg 2012). Shorter reads (36bp or less, typically used in ChIPseq rather than HTS) cannot be assigned to a single unique region of the genome with high confidence, and even doubling their length to 75bp results in 10-20% of the genome still missing a unique read asignment (Derrien et al. 2012). While those results are looking at single reads even paired reads, which both increase the single read length and provide additional uniqueness limitations by requireing that the mate read maps within a given distance, may not have a unique match in the genome. Even paired 100bp reads may map to 140 different locations in the human genome (Hach et al. 2010).

Both the amplification issues and the multiple mapping problem may be overcome in future advances in sequencing technology leading to longer reads. Currently there are three technologies aiming to generate these long reads: the SMRT sequencing platform from Pacific Biosciences claims to generate reads that average 14kb in length; Oxford Nanopore Technologies nanopore technology claims to generate 100kb reads; and Illumina has a “synthetic” long-read platform to generate fragments from 10kb sequences that are then computationally stitched together to create one long read (R. Li et al. 2015). Both Oxford Nanopore and Pacific Biosciences create their long read from a single molecule, removing the need for amplification and therefore potentially improving the uniformity of coverage. However, long-reads from single molecules suffer from a higher read error rate than current short-read platforms (0.1-1%) ranging from 4% as reported by Oxford Nanopore (not independently tested yet), to 14% for the Pacific Biosciences SMRT platform (Carneiro et al. 2012). While the raw read error rates make these technologies less suitable for whole genome sequencing on large genomes (e.g. human, mouse), they are likely to be a good tool for rapidly sequencing small viral or bacterial genomes, highly repetitive regions of the genome (e.g. centromeres and telomeres), or for targeted validation of larger structural variants. Currently this is all speculative as these

technologies are still new and continue to evolve and improve.

The third technology from Illumina, called TruSeq Synthetic Long-Read, uses existing instruments for generating the reads and alters the method. A new library preparation workflow cuts DNA into 6-10kb fragments and sequences these in parallel with the usual short-read workflow. The synthetic long-reads are then generated by local assembly of the longer fragments. This workflow has already demonstrated utility in sequencing repeat regions with significantly lower per-read error rates than single-molecule platforms. It should be noted that the biases due to amplification are still present in this technology and affect the local assembly of synthetic long-reads (R. Li et al. 2015).

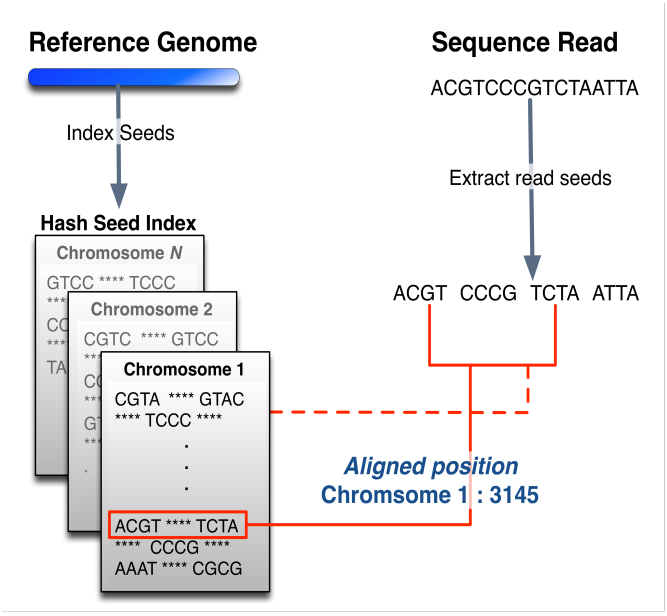
Long-read sequencing technologies offer significant promise for resolving current issues in sequencing, including the potential to improve structural variant sequencing in cancer genomes. However, as has been noted these technologies are new and currently untested in these areas and in the past several years thousands of cancer genomes have been sequenced on short-read platforms and made available to the community through the International Cancer Genome Consortium (ICGC) and TCGA.

1.3.2 Read Alignment

Completion of the Human Genome Project provided the first assembled reference sequence. That reference enabled the next leap forward in genomics that came with HTS, this time in the computational time required to map a new individual of the same species. The first genome assembly for a species (without a close relative species available) is generated through a *de novo* assembly process which overlaps individual reads into longer contigs, then generates scaffolds by overlapping the contigs (Baker 2012). Depending on the length of the genome this process can take hours to weeks (mouse/human genomes) on a computer with adequate memory and processing capabilities.

Once an assembly is available

Figure 7 Seed-and-Extend Search

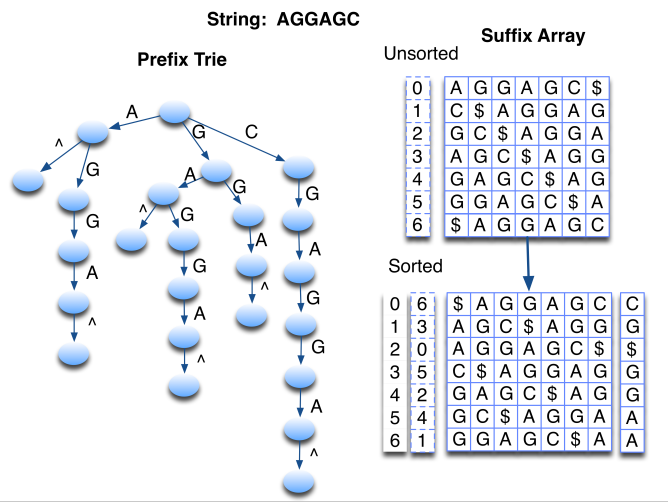


Seed-and-extend or hash search breaks both the reference index and the query string into seed kmers. The index is then searched with a query seed, when a match is found the string is extended, allowing for *N* gaps.

subsequent individuals can be sequenced and aligned in a fraction of the time depending on the genome size, coverage, and alignment algorithm used. The general process, known as *reference mapping* involves taking the single or pair-end reads output from an instrument and finding the locations in the assembled reference sequence that most closely match the read(s). This process is distinct from simply searching for motifs in a database (e.g. BLAST) or pairwise alignment between two different sequences (e.g. Clustal) due largely to the scale of the data. A single human genome sequence with 30X coverage can generate more than 450 million reads at 100bp lengths. Mapping that volume of data, allowing for mismatches due to SNVs or errors required more efficient algorithms. Two primary methods have been developed to address this problem (H. Li and Homer 2010):

- *Hash search.* The sequence (either the read or the whole genome) is hashed into *kmers* of a minimum length that is greater than 1 for efficient search. If the read was hashed the reference genome is scanned for matches, if it is the genome the read is compared to the hashed reference. In both cases when a match is found the *kmer* is extended until a mismatch occurs. This is also referred to as the *seed-and-extend* (see Figure 7) method and is used by MAQ (H. Li, Ruan, and Durbin 2008), and Novoalign (http://www.novocraft.com).

Figure 8 Burrows-Wheeler Transform



BWT builds the reference index as a suffix array/prefix trie. The query string is then compared against the prefixes to find the maximal paths. Image based on (H. Li and Homer 2010).

- *Burrows-Wheeler Transform.* This algorithm uses suffix/prefix trie data structures where the string is represented as an array (list) of strings generated from the initial string by rotating the suffix then sorting by the symbols that start each member of the list (see Figure 8). The corresponding prefix trie (a graph data structure) represents the reverse of the suffix array and can be traversed to enable fast searching. This method is used by BWA (H. Li and Durbin 2009) and Bowtie (Langmead, Trapnell, et al. 2009) to create a searchable trie from the reference genome.

The ultimate goal of sequence read alignment is not simply to report the sequence, it is to identify the differences between species, variations within a population, or heritable mutations in a family. This means understanding and accepting the limitations of both current sequencing technologies and alignment methodologies is necessary.

1.3.3 Structural Variant Identification

Identifying structural variants (SV) in short-read sequences from tumor samples is not a simple task. Breakage and recombination of the chromosome appears to happen at fragile locations, directly altering the sequence by combining fragments from other chromosomes or from other locations within the same chromosome. Read-pairs generated from this region of the chromosome could span the breakpoint if it fell within the gap of the read-pair, or "split" the read so that the beginning and end of a single read aligns to different chromosomal locations. Computational limitations in the alignment algorithms make correct or unambiguous alignment of these read-pairs difficult. In fact many of these read-pairs may not be aligned to the reference at all (Schbath et al. 2012; Ruffalo, LaFramboise, and Koyutürk 2011). The rate of read-pair alignment failure in cancer genomes can vary from nearly normal at about 3% of the reads failing to align, up to as many as 40% of the reads being unmapped.

The identification of structural variants in tumor samples is complicated further as there is a high degree of genomic heterogeneity (see Section 1.2). Depending on the specific type and grade of tumor, a sample taken from a solid tumor is likely to include multiple sub-clonal cellular populations that do not share all of the same variations (Greaves and Maley 2012). The result in sequencing and alignment from such samples is a low frequency of read-pairs supporting a single SV position, or those reads failing to align altogether and thus unavailable in the process of SV identification. Multiple approaches to identifying SVs from such samples have been developed due to these ambiguities.

Reference Based

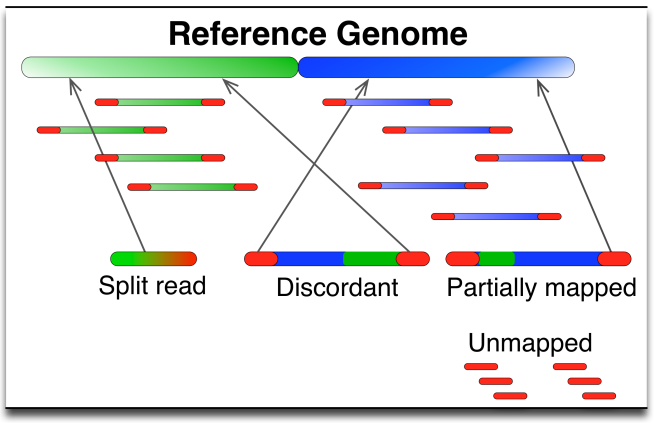
The most commonly used methods for identifying SVs in short-read data currently rely on the alignment locations reported by the alignment algorithms. These methods can be classified as *Reference Based*, as they entirely rely on the alignment and reporting of reads to locations relative to the reference genome. Due to the fact that the reference genome will not be representative of an instable genome that typifies cancer (e.g. due to one or more chromothriptic events), the resulting alignment will include many more reads that are either (see Figure 9):

- *Discordant* This is generally defined as: read-pairs for which the alignment algorithm mapped each read of a pair to a different chromosome; where the insert distance between the read-pairs is larger than would be expected (e.g. greater than 4 s.d. from the mean based on the read library); or where the orientation of the read-pairs is incorrect (relative to the specific read technology).

- *Split-reads* Where a proportion of a single read may align to the reference, while the remainder is "soft clipped" or unaligned, this is indicated in the alignment BAM file with the CIGAR value. These reads are often considered to be spanning a breakpoint location directly.

- *Partially mapped* These read-pairs are not fully aligned, as one of the pair will report alignment to a location on the reference while the other read does not align at all.

Figure 9 Inconsistent Read Alignments



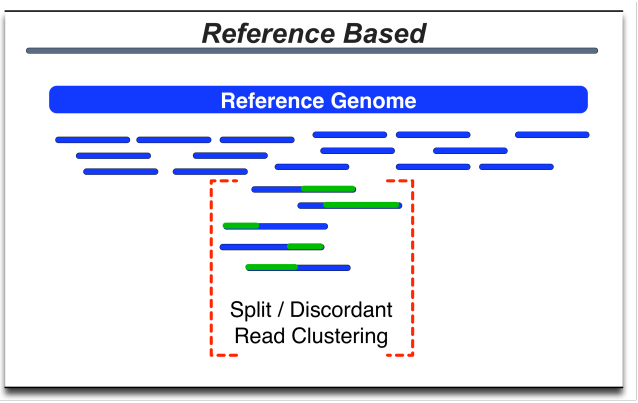
Alignment of reads which may be a result of structural variation.

A fourth category of reads is present within all sequence alignment samples: read-pairs where neither read was aligned to the reference. In sequences from non-tumor samples, or "normal" tissue, these unmapped reads typically account for 2% or less of all reads. By contrast, in tumor samples up to 40% of reads may be entirely unmapped. These reads are lost to all SV and SNV detection methods that rely on reference based alignment.

The aligned (discordant, partial, and split) reads can be used to infer the existence and position of a breakpoint (see Figure 10) through clustering or windowing strategies (Medvedev, Stanciu, and Brudno 2009). Discordant reads are used by tools including BreakDancer (Chen et al. 2009) and Pindel (Ye et al. 2009) to identify breakpoints through clustering the reads by the locations the reads aligned to. While tools such as PRISM (Jiang, Wang, and Brudno 2012), DELLY (Rausch et al. 2012) and SoftSearch (Hart et al. 2013) cluster split-reads or partially mapped reads, typically to identify smaller variants (e.g. deletions, insertions). As these methods are usually limited by the size of the variants they can detect consensus approaches such as SVMerge (Wong et al. 2010) are often used to increase detection across all variant classes.

The disadvantages with this approach are directly related to the reliance on the reference genome and algorithmic constraints on the alignment itself. While hashing and suffix/prefix tries can be used to perform exhaustive searching for read matching, it is infeasible due to the computational cost of performing an exhaustive search for inexact matches as is necessary for short reads. In fact, most of the algorithms are not set up to allow for exhaustive search, particularly with regards to repetitive mappings, relying instead on heuristics to allow them to return the "best mapping" or a randomly selected "best" where there is not a single unique alignment (Treangen and Salzberg 2012). This solution is reasonable where the reference genome and the sample are reasonably closely aligned. With billions of reads most repetitive regions will be covered by the random selection of non-unique mappings, and reporting of the best mapped reads including *n* mismatches will provide high quality mappings for variant detection (H. Li, Ruan, and Durbin 2008). However, in tumor genomes with high rates of both small (e.g. SNP, indel) and large (e.g. translocation, inversions, deletions) scale variations the rate of alignment can drop significantly as the difference between the read and the reference is too large. The consequence for reference based methods of structural variant detection is fewer reads that are usable for breakpoint detection and many breakpoints that are entirely undetectable. Due to these limitations methods that rely less on a reference genome are being developed.

Figure 10 Reference Based Methods



Reference based approaches first align reads to the reference, then clustered using a variety of measures to identify SVs assuming that the alignment is correctly and uniquely reported.

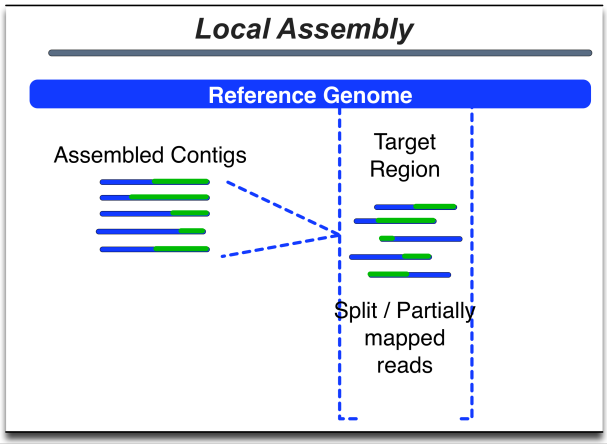
Local Assembly

One possible approach that does not rely on a reference would be a complete *de novo* alignment of tumor genomes. However, assembly of genomes *de novo* is complex in normal human samples (Schatz, Delcher, and Salzberg 2010) due to their size and repetitive sequences. Aneuploidy, chromothripsis, and sample heterogeneity increase the difficulty in tumor samples and make *de novo* assembly of the entire sequence impractical for general use. Instead *de novo* methods can be employed in smaller, targeted regions to perform local assembly in a similar manner to resequencing studies where specific

regions are sequenced and assembled rather than entire genomes.

In general *de novo* methods use kmers of sequence reads to create various data structures including *De Bruijn* graphs or overlapping graphs in order to build longer contigs of a sequenced region. A new approach using kmers in targeted regions of tumor genomes called BreaKmer (Abo et al. 2014) employs this general approach in tumor genomes without first resequencing regions. Instead reads that are split or partially aligned within a targeted region are reassembled into contigs using overlapping kmer seeds from both the reads and target region in the reference (see Figure 11). These contigs are realigned within the target regions using BLAT and classified into variant types (e.g. inversions, indels, deletions). Local assembly of contigs from regions that are highly structurally variant, and therefore have high rates of read misalignment, overcomes one of the most significant issues in short-read sequencing – that such short reads are unlikely to be uniquely aligned where the genome is structurally altered. This approach is not currently appropriate for SV identification in a genome-wide context as it involves *de novo* assembly on a local scale, however it is highly useful for resequencing experiments and identification within targeted regions.

Figure 11 Local Assembly Methods



Local assembly takes misaligned reads to build longer contigs of the putative breakpoint regions followed by realignment against a reference.

Reference Free

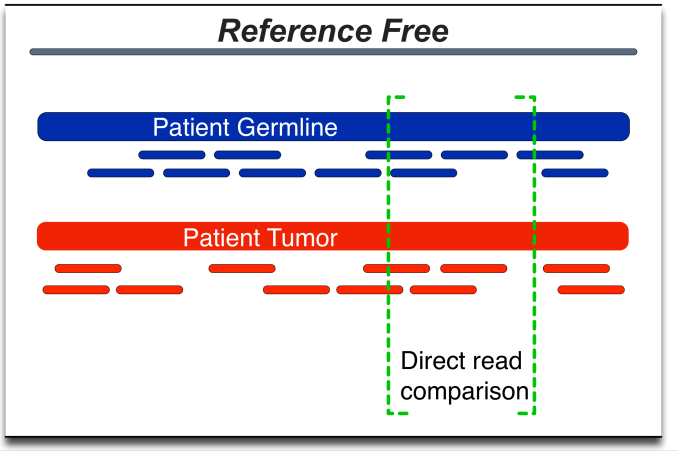
Another approach that has been taken is to avoid the reference altogether and directly analyze the sequence reads output by the instrument. In this approach the reads do not need to first be aligned, as positional information is not taken into account (see Figure 12). This approach is new and therefore the methods of analyzing the reads vary widely.

The authors of the CommonLaw tool (Hormozdiari et al. 2011) assume that structural variants can be detected with higher accuracy by simultaneously analyzing multiple related genomes using the reference genome as an intermediary. The true variants are assumed to be discoverable by comparing the patient genomes directly and this is shown with small structural variants (<1kb) in YRI genomes (The 1000 Genomes Project Consortium 2010) and a family trio. While the authors state that it should work in tumor genomes, they do not try it and it is unclear how direct comparison will work with the

complexity of tumor samples.

A more recent approach called SMuFin (Moncunill et al. 2014) was developed specifically for tumor samples and directly compares reads between two whole genome sequences without alignment. In this case the two samples are the normal and tumor pair from a single patient. It is expected that reads from both samples will be highly similar, and that mutations can be identified by grouping the reads into a tree structure (based on a generalized suffix array) that branches where mutations occur. Breakpoints can be identified in the branches as well as SNVs, and local assembly performed on the reads within those branches. Both SMuFin and CommonLaw identified small structural variants with greater accuracy than the methods that rely on the reference genome. Furthermore, the analysis performed with SMuFin showed that there is significantly more complex large-scale structural variation in tumor samples than has been previously reported. Due to the massive amount of computation involved (e.g. comparing billions against billions of reads) these methods are computationally very expensive even when massively parallel systems are used (e.g. SMuFin requires tens of machines to analyze a tumor/normal 30x coverage genome pair) making this approach difficult to adopt in high-throughput settings where many patients are being sequenced and analyzed.

Figure 12 Reference Free Methods



Directly comparing reads from two genomes (e.g. tumor and normal) enables variation identification without use of any reference.

1.4 High Performance Computing for Genome Analysis

As the scale of biological data has reached “big data” proportions the difficulty of handling that data has increased as well. Sequencing technologies have continually improved over the last 10 years, with constantly increasing coverage at decreasing costs. Computer performance has not been able to keep up with this explosion of data. This has led to the development of many pipelines and tools that use parallel frameworks to break down the computation into discrete chunks that can be recombined later. No matter which SV detection system is used there is a growing need to ensure that it will scale, due to the rise in the number of genomes and the growing interest in personalized medicine.

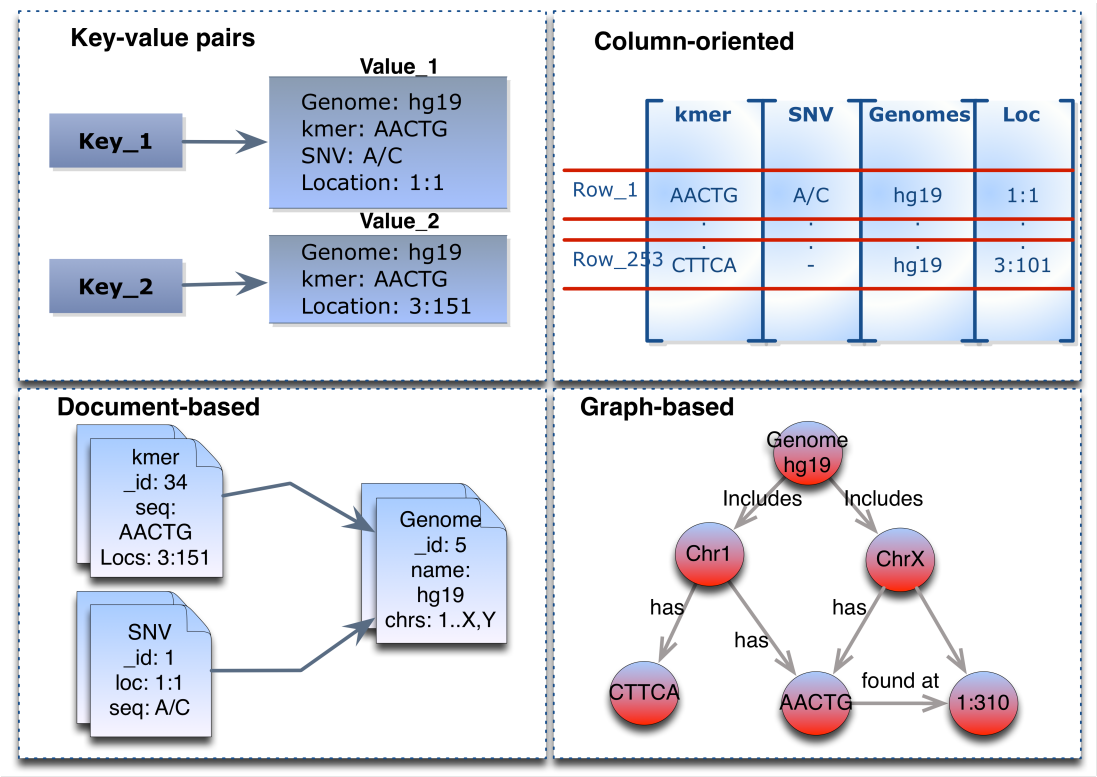
1.4.1 Distributed Data Access

The size of genomics data is an issue for any sort of analysis that makes direct use of the sequence itself. In generating new references that are mutated in any way (e.g. SNVs or SVs) direct access to a section of sequence in a specific location requires filtering through thousands or millions of base pairs in FASTA files. An indexing strategy can simplify this if the primary sequence is the only one being accessed, however in looking to generate related trees of sequences to simulate generations of cellular mutations (as in this thesis) it is clear that a highly distributed database solution is necessary to track and access mutated sequences.

A similar problem is faced in searching through reads from a sequencing sample. In a single sample millions or billions of reads are generated. Many of these will also be closely related due to the depth of coverage at each base pair. This is multiplied when searching across multiple samples as most “normal” reads are closely related between individuals. A distributed database solution would enable storage of the reads directly, compressing highly similar reads, as well as metadata about each read. However, in both cases the data needs to be flexibly stored as the requirements for analyzing the raw data or specific metadata are constantly changing. A traditional relational database (RDBMS) does not provide both the scalability and the flexibility required for storing sequences from large-scale genomics data.

While a variety of database solutions are available that operate within a distributed system, including both RDBMS and NoSQL implementations, NoSQL databases are more commonly used for massive data where the structure may be unclear or frequently changing. Distributed databases do not aim for transactional consistency (e.g. ensuring that every step of a process is consistent and isolated from all other operations) at all times as a RDBMS does; instead they provide fully redundant storage by distributing the data across the system of nodes (machines). NoSQL databases do not require prior data structuring or the creation of a data model that is encoded in relational tables, instead database design is meant to directly model the application use and is highly dependent on the database category (e.g. graph-based information may not be easily encoded into a document format). There are four major categories of NoSQL databases based on how they store data (see Figure 13):

Figure 13 Distributed Databases



NoSQL databases are encode the relationships between different pieces of data directly.

- *key-value pairs* where hash tables containing a unique key are paired with a value that may be simply strings of text or a data structure such as JSON (example: MemcacheDB)
- *column-oriented* where data is stored in sets of columns that are related and can be retrieved as a group based on a single key (example: HBase, Cassandra)
- *document based* stores data similar to a key-value pair, but the pairs are stored in a document encoded using JSON or XML and are ordered hierarchically in a tree structure (example: MongoDB, CouchDB)
- *graph based storage* encodes objects as nodes in a graph with the relationship between them represented in the edges (example: Neo4J)

Many different large-scale scientific applications have used these NoSQL solutions to provide data storage and access for unstructured or semi-structured data including particle physics data archives for CERN using Cassandra, and genomic variation data for the European Variation Archive using MongoDB. Each of these can also be connected to various distributed computing frameworks for rapid analysis (such as MapReduce based ones discussed in the next section).

1.4.2 MapReduce for Genomic Analysis

The most commonly used model for parallelizing biological data analysis has been MapReduce (Dean and Ghemawat 2008), as it is a simple method for using commodity or server hardware as clusters for computationally intensive tasks (T. Robinson et al. 2011; Lewis et al. 2012; Lewis et al. 2010). Each MapReduce job is broken down into three distributed and parallel steps:

- *Map* phase performs a user-defined computation on a segment of data; the output is a key/value pair that relates to the computation or data used.
- *Shuffle/Sort* phase takes data from the mappers based on the keys they output, the data is sorted and again emitted with key/value pairs
- *Reduce* phase processes the output data from the shuffle/sort phase and finally outputs it to a file system or database.

Each mapper runs on a single node, with a master node orchestrating both data redundancy and assigning data to each of the mapper nodes. A MapReduce analysis should then scale linearly to the number of available nodes, though this depends on how independent each Map or Reduce step is and on the location of the input data. The ideal usage of the system involves planning the mappers on nodes that are close to the data being analyzed as the less the data itself moves the faster the overall process will be.

In the context of genomic data this framework has been especially suitable for a variety of alignment problems. A single whole genome sequence can have billions of sequence reads. Each of these may be initially aligned independent of any other read pair and therefore need access to only the reference genome index. In MapReduce this process is broken down relatively simply: each *Mapper* runs an instance of the aligner and aligns a chunk of reads against the index; the aligned reads are *Shuffle/Sorted* so that all alignments to the same chromosome are together and sorted by location; and finally a *Reducer* can merge the alignments into a single BAM file. This is the basic workflow for the alignment portion of Crossbow (Langmead, Schatz, et al. 2009) using Bowtie, the Hadoop-BAM (Niemenmaa et al. 2012) library using BWA, and CloudBurst (Schatz 2009) a novel *seed-and-extend* aligner developed specifically within the MapReduce framework. The resulting aligned BAM file has lost no read information in this process. The alignment itself is simply sped up by enabling multiple independent read pairs to be aligned simultaneously.

Subsequent analyses, such as variant calling, have to be designed to consider the reads in context to surrounding chromosomal regions. Since these regions can still be broken into chunks the MapReduce paradigm continues to be useful here as well, with

each *Mapper* taking only a segment of data and calling variants based on the reference locations and particular model of the variant caller. This process has been highly effective and the Genome Analysis Toolkit (McKenna et al. 2010), one of the most commonly used genome informatics applications, was built to wrap the MapReduce process to make using the framework for developing common genomic analyses even simpler.

1.5 SUMMARY

New methods for variant detection are continually being developed. Many have focused on the small variants (e.g. SNV, indels, short repeats) that may be common in diseases that have far less genomic chaos or diversity than cancer. Others have focused on those structural variants that may be relatively easy to identify in cancer such as somatic copy number aberrations. All of these variations have been shown to be important to driving the progression of cancer however, due to the massive chromosomal rearrangements and breakages during a chromothripsis event variations ranging from large deletions, duplications, translocations and inversions are also important. Identifying these large-scale variants, particularly those that are copy-number invariant, in short-read sequencing data continues to be difficult.

Methods to identify and characterize these large-scale variants, such as the one discussed in this thesis, must be capable of: handling short-read data with an awareness for the limitations of the technologies and their impact on the reads themselves; scaling appropriately to both the size of single genomes (e.g. billions of reads) and ultimately to populations of genomes; accounting for the heterogeneous nature of cellular populations within a tumor sample and the resulting issues with low-coverage structural variants. The solutions to these issues as developed in this thesis are presented in the Methods and Results chapters.

CHAPTER 2 SCOPE & AIMS OF THESIS

The high complexity and diversity of cancers presents a significant challenge to researchers in understanding the genomic basis for development and progression of the disease. Individual genomic variation and instrument sequencing biases complicate data analysis, but the primary difficulty lies in the sampling and subsequent sequencing methods. Sample collection typically results in a mixed population of cells that includes adjacent tissue, normal tissue and discreet subpopulations of tumor cells. Such subpopulations may have large genomic differences that can drive clonal expansions, drug resistance response, or metastatic disease. Whole genome sequencing (WGS) of these tissues is typically done without applying any separation to the tissues and cells.

Thus the major problem for cancer genomics (using short-read sequencing) is that this heterogeneous tumor sample is treated as a homogenous sample due to the technological difficulty in sequencing and analyzing such mixed samples. Downstream analysis of the sequences are then complicated by the limitations of alignment algorithms that rely on reference that does not reflect the underlying genome structure. Ultimately these analyses are missing the evidence of large structural variants that may appear at low frequency in the sample.

Due to these difficulties multiple methods have already been proposed to identify large-scale structural variants. These have been classified into three different approaches (e.g. *Reference Based, Local Assembly, and Reference Free*) and are described in the first chapter (section 1.3). Each of these approaches has their own advantages and weaknesses, and it is clear that there are still many variants that are not being identified.

This thesis proposes an approach to identify large-scale structural variants with these issues in mind by using available information about cancer breakpoints to generate hundreds of synthetic references for simultaneous alignment and analysis in heterogeneous samples.

2.1 THESIS AIMS

Aim 1: Create in silico references by providing the means to generate a large number of synthetic sequence references that reflect possible genomic rearrangements. This involves generating the background information on cancer-related breakpoints by analyzing karyotypes to determine the distribution and frequency of large-scale structural aberrations. The resulting background will then be used to generate new references that model probable structural variations that can be used for subsequent alignment.

Aim 2: Multiple reference alignment and structural variant detection by aligning the references generated in Aim 1 against reads generated from patient samples. The result will be hundreds of alignments that will then be evaluated for representation of large-scale structural variation by the aligned patient reads. A scoring metric for alignments to references that model SVs represented in the patient data will be developed by taking into account read quality, sequencing limitations, and the selected alignment algorithm. Additionally, appropriate high-performance computing tools will be used, as multiple-alignment will be a computationally intensive task.

Aim 3: Validate the method using simulated sequence data, and test in patient datasets available from the Cancer Genome Atlas project (The Cancer Genome Atlas 2008; The Cancer Genome Atlas Network 2012). This requires that appropriate simulated data be generated to model breakpoints in sequences at varying coverage levels. This aim will use available read simulation tools to generate sequences that include breakpoints, followed by detection and scoring using the methods developed in Aim 2. Finally, TCGA patients representing different cancer types will be analysed for probable SVs.

2.2 Originality

Multiple-reference alignment (called *MultiSieve* in this thesis) is a method for evaluating genomes that may contain large-scale structural rearrangements compared to a 'normal' genome. Previous approaches have used multiple closely related genomes (e.g. family trios or tumor/normal paired samples) to separate the process of variant identification from location alignment by directly comparing reads first. These methods showed that the standard reference-based approaches were missing identifications. As chromothriptic events and large-scale genomic rearrangement are associated with later stage cancers they may be important to understanding drug resistance and metastasis, while identifying these structural changes in early stage cancer may be valuable for prognosis or drug targeting. That these rearrangements are likely to be found in small subpopulations of the tumor that have not yet become invasive or been through a positive selection and clonal expansion event make them both more difficult to identify, and more valuable to find.

This project aims to develop a method that is more sensitive to the heterogeneity present in tumor samples through the use of prior knowledge to generate multiple references and high-performance computing to rapidly align and analyze the sample sequences.

CHAPTER 3 MATERIALS & METHODS

Current alignment and variation identification methods require a single whole genome reference to map sequence reads and subsequently identify variant locations. The reference used is generally composed of the 22 autosomal chromosomes, both of the sex chromosomes (regardless of the sample sex), the mitochondrial chromosome sequence, and often viral genomes that may be expected in the sample (e.g. HPV, EBV). These are provided as a single large reference regardless of the patient being analyzed. These large reference alignments have the underlying assumption that the reference used is highly homologous to the sample being sequenced. In cancer, due to structural variations and high mutation rates, this assumption does not hold.

To reliably detect cancer structural variations an alternative approach, which does not rely entirely on reference based alignment, is needed. This thesis proposes one such method, which uses *in silico* generated references: the MultiSieve method.

Instead of only aligning against one large reference sequence, in silico reference alignment uses a set of small references that model likely structural variations. The poorly aligned or unaligned reads from the cancer sample are then aligned against all of the small references. If a structural variation occurs in the sample that has been modeled by one of the small references then the alignment will be highly ranked as compared to non-representative reference models. Such an approach improves upon existing methods, as it is able to detect structural variations that single reference based alignment systems simply do not search for. The reason they do not search for them is due to the non-exhaustive alignment strategy used by these reference based aligners (and a lack of analyses to support exhaustive alignment reporting), as an exhaustive search on a human genome would require years of computational time and is not practical on todays hardware.

The in silico reference technique depends upon the *de novo* generation of a large number of small references. However, a random generation of possible variations for each sample to be analyzed would be extremely time consuming computationally, and is unnecessary, as some information about variations is already known. Therefore this method relies on the development of a knowledgebase to inform reference models of structural variation that may result from complex chromothriptic events as well as smaller mutations.

Section 3.1 (Reference Selection) describes the approach used to analyze the variations required for an informed generation of multiple references. Section 3.2 describes the methods used to align the model references against patient reads and

score them for structural variation inclusion. In this section the criteria applied to evaluate patient alignments and the probable structural variations are defined. Section 3.3 describes the HPC methods required for the computational needs of such a method.

3.1 Reference Selection

While simulating mutations and structural changes to the genome can be achieved through random perturbations to the sequence of the standard reference genome (e.g. GRCh37), this has limited utility with respect to aligning patient reads due to the computational time required to align and analyze each set of reads. However, clinical data exists which describes structural alterations at a karyotype level (e.g. observed through microscopic imaging). Using these data to optimize the generation of simulated references provides cancer-specific information to guide the simulation, and limits an otherwise "very large" search problem.

A reference in the context of genome alignment is a FASTA sequence file that is indexed by the alignment algorithm for rapid search. In the case of the human genome the indexed FASTA file will most often include all 22 autosomal chromosomes and both sex chromosomes (also typically mitochondrial genome and putative integrated viral genomes). The reference itself can therefore be altered to provide a more accurate comparison to the patient genome.

In the case of structural variations there are no references that provide for the range of large-scale (involving several kb of sequence) alterations. A reference that includes structural variation would require that the sequences of two or more chromosomes be segmented, shuffled, removed, duplicated, and sometimes inverted. The entire sequence of each FASTA chromosome sequence would include aberrations that result from breakpoints and fusions with other chromosomes (representing the result of chromothriptic or translocation events).

To optimize the construction of these references they were generated synthetically through application of known mutation rates for normal and/or cancer genomes (Killcoyne and del Sol 2014), and by using available data on cancer structural variations observed in spectral (SKY) or FISH karyotyping (Schröck et al. 1996; Speicher, Gwyn Ballard, and Ward 1996). The method for identifying the mutation rates, and putative breakpoints at the karyotype level is discussed in section 3.1.1.

To ensure that the system could run effectively (in hours rather than days) on patient samples a selection algorithm was additionally implemented. This selection algorithm provided the means to optimize the choice of in silico references, and is described in section 3.1.2.

3.1.1 Karyotype and Mutation Analysis

Two types of mutation analysis and artificial genome generation systems were implemented to support this work:

1) A karyotype level detection system, used to store information about known structural variations that occur in cancer
2) An indel level mutation system, used to model small-scale nucleotide level mutations that occur in cancer.

Both of these systems used public data about known mutations. In the main MultiSieve alignment system only the first (karyotype level reference) was used, as this provided the large-scale structural variation information that was used to generate the required references.

The second (indel level) system was used primarily for benchmarking the computational needs as it allowed for the rapid generation of a large number of simulated "cancer like" genomes. These artificial genomes were required as they included known characteristics and avoided any issues associated with patient anonymity, security, or non-transfer agreements.

Table 3 Publicly Available Karyotype Data

Source	Date Downloaded	# Karyotypes	Patients	Cell lines
Mitelman Database (F Mitelman, Johansson, and Mertens 2015)	2012-11-26	99,764	✔	
NCBI SKY-FISH Database ("NCI and NCBI's SKY/M-FISH and CGH Database" 2012)	2012-11-12	325	✔	✔
University of Cambridge CGP (Edwards 2012)	2012-10-22	84		✔
NCI Fredrick National Laboratory ("NCI Fredrick National Laboratory Cell Line Drug Discovery Panel" 2012)	2013-01-16	67		✔

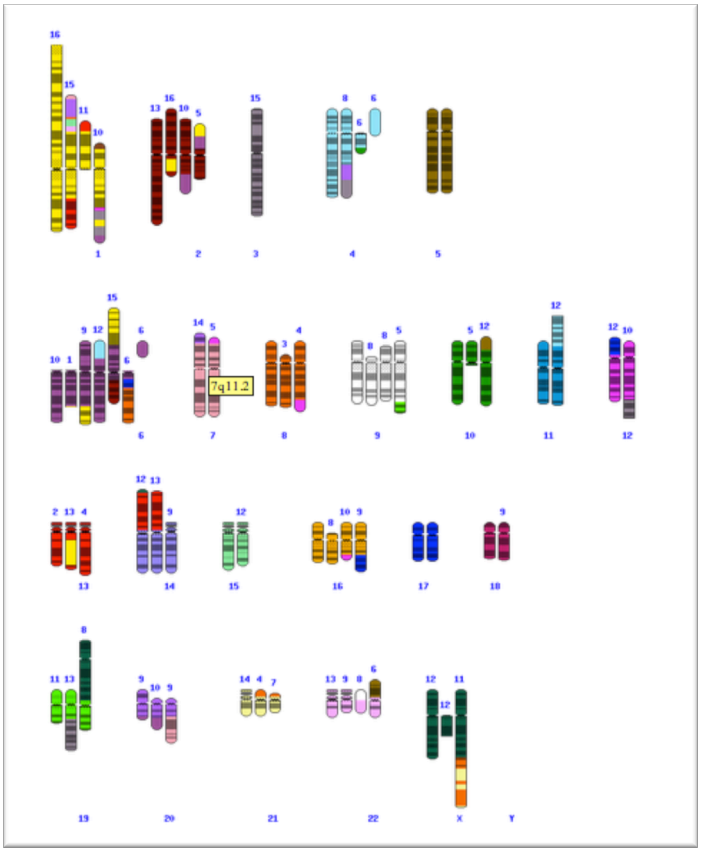
Publicly available karyotypes from cell lines and patients were downloaded from various sources. The largest source of patient-derived karyotypes is from the Mitelman Database (Felix Mitelman, Johansson, and Mertens 2007).

Karyotype Level

To model large-scale structural variants that typify cancer, a mixture of both patient-derived (see Figure 14) and cell-line derived karyotypes was downloaded for analysis. The majority of these, nearly 100,000, came from the Cancer Genome Anatomy Project (CGAP) under the Mitelman Database of Chromosome Aberrations and Gene Fusions in

Cancer (F Mitelman, Johansson, and Mertens 2015), a curated database of chromosomal aberrations in tumors. The remainder, about 500 karyotypes, came from three other sources including the NCBI SKY-FISH/CGH Database, the Cancer Genome Program at Cambridge University, and the NCI Fredrick National Laboratory CellLine NCI60 Drug Discovery Panel (see Table 3).

Figure 14 Karyotype from Pancreatic Cancer Patient



Example visual representation of a karyotype from a male patient with metastatic pancreatic cancer (Barenboim-Stapleton et al. 2005). This karyotype displays multiple large-scale structural variations as well as aneuploidy. Downloaded from the NCBI SKY/M-FISH & CGH Database ("NCI and NCBI's SKY/M-FISH and CGH Database" 2012)

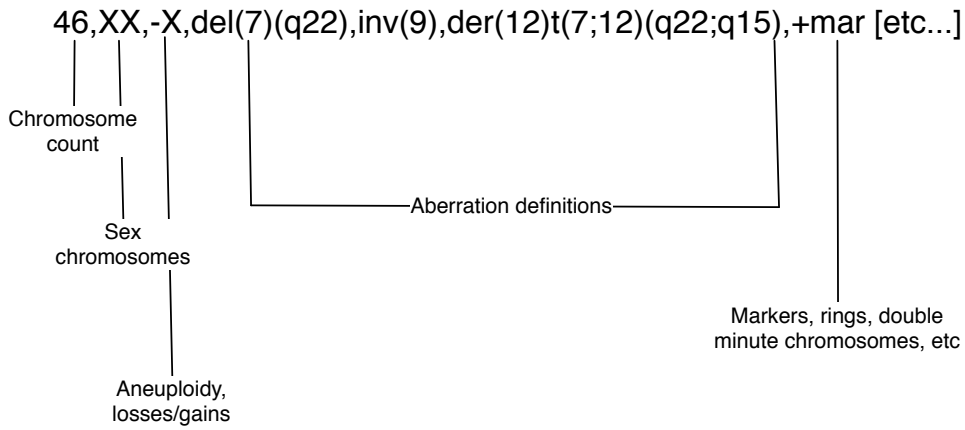
This background data encompassed 227 cancer types. When dealing with this data, inconsistencies in annotations, problems with metadata, data format inaccuracies, and biases in the data had to be addressed.

The cancer types were not always useful descriptions of an ontological type, for instance "heavy chain disease", "heart" or simply "leukemia" were not uncommon disease labels. On the other extreme were descriptions (primarily from the Mitelman Database) that were overly specific: "Atypical lipomatous tumor/atypical lipoma/well-differentiated liposarcoma" or "Acute myeloblastic leukemia without maturation (FAB type M1)". As closely as possible these descriptions were mapped to the US National Cancer Institute

(NCI) cancer types.

Based on these descriptions it was determined that 66% of all karyotypes found in the Mitelman dataset were blood-related cancers, primarily leukemias. This was important to the breakpoint analysis as leukemia is known to accumulate structural variation at a high rate, and there are known prognoses associated with various recurrent variants (Welch et al. 2012). Thus breakpoint information reporting is biased towards variants that are common in leukemia.

Figure 15 ISCN Karyotype Definitions



ISCN karyotype definition contains several scales information from genome-wide (e.g. chromosome count, sex) to individual aberrations. When karyotypes are submitted to the various databases, or published in papers, there is no check that these conform to the ISCN. Each that failed to parse correctly had to be manually curated before breakpoint information could be determined.

These karyotypes were provided in text form using the International System for Human Cytogenetic Nomenclature (ISCN) standard format (see Figure 15). This format required significant cleaning before an analysis of breakpoints was undertaken. Various parsers have been developed, however as the ISCN standards have been frequently updated and karyotypes themselves are rarely used in a high-throughput environment an updated parser had to be developed in order to generate the necessary information.

Most karyotypes failed to follow the ISCN standards in one or more definitions and required manual cleaning, or interpretation. Small errors could have resulted in losing all breakpoint information from a chromosomal aberration such as: incorrect notation of a segmental deletion by using '-12' (the standard format for chromosome loss) instead of 'del(12)(q14)'; incomplete information for derivative chromosomes (e.g. 'der(7;12)' with no bands); or aberration definitions that included a '?' indicating that the specific aberration could not be determined. Other errors required simple curation, such as missing semi-colons or incorrectly used definitions (e.g. 'add(7;14)(q12;p14)' probably indicated a derivative chromosome with a translocation). The parser developed to read these

karyotypes was intentionally conservative in the interpretation of aberrations due to the need for well-defined breakpoints.

Breakpoints in a karyotype were defined as any aberration which was expressed in a format that included both a chromosome and a band (e.g. 'del(12)(q14)' defines a breakpoint while 'i(9)' does not). Exceptions to this were aberrations that were only partially defined (e.g. one breakpoint can be determined, but the '?' leaves the second breakpoint unknown t(1;12)(p32;q?)) and ring chromosomes (e.g. 'r(11)(p15q25)') due to inconsistent band definitions. The result was background data on 30,558 aberrations, not including whole chromosomal gains or losses, in all chromosomes and including all 320 major cytogenetic bands.

The cleaned and parsed karyotypes were subsequently analyzed for aberration information. Aberrations were classified by ISDN standard types including deletion, duplication, inversion, and translocation. Across all karyotypes 14 aberration types were used, with one additional class for aberrations whose classification was unclear and could not be resolved manually. Aberrations were then further broken down into breakpoints where appropriate (not all aberrations indicate a breakage as in the case of aneuploidy or double minute chromosomes). Breakpoint frequency and chromosomal instability was determined based on this final parsed information, and output was generated as simple text files or database tables for use by the selection algorithm.

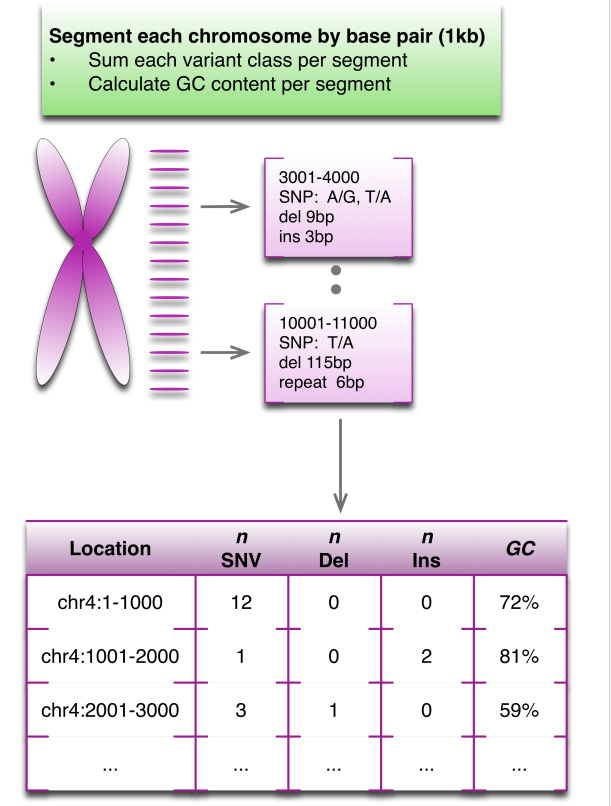
This background data was then stored in a specially designed database, and used for the basis of generating the in silico references required by the MultiSieve system.

INDEL LEVEL

To enable the generation of in silico cancer genomes (Killcoyne and del Sol 2014), data about small scale mutations was also collected and collated. Human genomic variation has been evaluated through various large projects that offer a wealth of public datasets characterizing population-specific variation. These include data on the normal range of variation from the 1000Genomes (The 1000 Genomes Project Consortium 2010), HapMap (The International HapMap Consortium 2003), and allelic frequencies in a population from dbSNP (Kitts and Sherry 2002) to sources that provide disease specific variation such as the Catalogue of Somatic Mutations in Cancer (COSMIC) (Bamford et al. 2004) for variations identified in cancer.

These public datasets primarily identify small variants (e.g. less than 1kb in length) including single nucleotide polymorphisms (SNP), indels, substitutions, and tandem repeats. Analyzing variation frequency required first identifying those data sources that provided the 'normal' (or non-cancer) level of variation across the genome as the somatic mutation rate in tumors (Ding et al. 2010) can be significantly higher than in normal tissue.

Figure 16 Segment and variant analysis



Each chromosome is segmented into 1kb fragments. A profile of small variants per fragment is determined and the GC content recorded.

Characterizing the normal mutation frequency involved mining variation data that was identified or validated by 1000Genomes and HapMap in the Ensembl datasets (Y. Chen et al. 2010). These data were provided in the form of GVF (Genome Variant Files), which reports the chromosomal location, variation type, and validation dataset for each variation based on reference genome GRCh37 (also referred to as hg19). To profile the mutations a window-based approach was used, where mutation rates and classes were collected in 1kb windows (or bins). This bin size was selected as: no small variant was reported to be larger than this; it was small enough to offer a reasonable profile of sequences; and was a reasonable size for subsequent computation.

In each 1kb segment the individual variants were characterized for the type or class (e.g. SNP, indel, substitution) and size (e.g. 1bp, 8bp, 234bp) of each. Several profiles for each 1kb segment were based on this and included the count for each class of variants, the size for each class other than SNPs, and the specific base-pair mutation in the case of SNPs.

These profiles were analyzed to determine the frequency of each class per fragment across the genome. For instance in the first segment of chromosome 4 there may be 12 SNPs and no other variants, while in the third segment there may be 4 SNPs, 1 deletion of 8bp, and a substitution of 3bp. Finally, structural elements of the sequence fragment were analyzed to identify directly observable elements that correlate with the variant frequencies (see Figure 16). These include:

- Identifying the incidence of coding/non-coding regions contained within the segment and correlating these to the different classes and total number of each variant class within a fragment. This included separating out exonic versus intronic regions as well since previous work has shown a link between active sites and SNV mutations.

- Overlaying predicted CpG methylation sites (Das et al. 2006) on each segment and testing for correlations of each variant class. SNVs in particular have been linked to methylation regions in CpG islands and in tumor sequences these are linked to particular mutator phenotypes. It was therefore possible that this phenotype starts in 'normal' somatic mutation.
- Determining the GC content for each segment (simple count of GCs per 1kb segment) and test for correlations between total, high, or low GC content. GC content is related to, but not always found with, CpG islands and is also related to evolutionary mutation rates. This could therefore be related to the overall somatic mutation rate in normal as well as tumor tissues.

Each of these analyses involved integrating other data types per segment including the coding regions as defined by Ensembl Genes or Transcrips, CpG methylated and unmethylated regions as well as CpG islands as predicted by Das et al. 2006. The same frequency analysis was performed on data from small cancer variations as reported in the COSMIC and the Database of Genomic Variants Archive (DGVa).

This information about mutation rate frequency was stored in a specially designed database and was used by a MapReduce genome generation system discussed in section 4.5 in Results.

3.1.2 Reference Selection Optimization

There are 320 major cytogenetic bands within the human genome (Kirsch et al. 2000), and each of these is involved in at least one aberration reported within the karyotypes described. A pairwise combination of each of the bands to create simulated references results in 51,040 possible combinations. Ideally each of these could be tested against a genome, however there are current computational limitations for this approach:

- *Limited hard disc space.* The index for all simulated references combinations requires 2.5 TB of hard disc space, and the subsequent alignment BAM files for a small number of reads (1.9 million) from a single genome would require more than 30 TB. Even keeping these BAM files temporarily would require a large dedicated storage server to enable the analysis of multiple genomes simultaneously.
- *Computational time.* Aligning a small number of reads against a bank of smaller references is an ideal situation for parallelization, however each single alignment (e.g. bwa mem –a –t 12) plus analysis computation required over an hour on a single node (with four cores), and there are over 51K possible alignments. If all of a representative 2100 computing cores cluster (the total size of the Luxembourg University HPC facility) were used solely for one run on a single patient (under

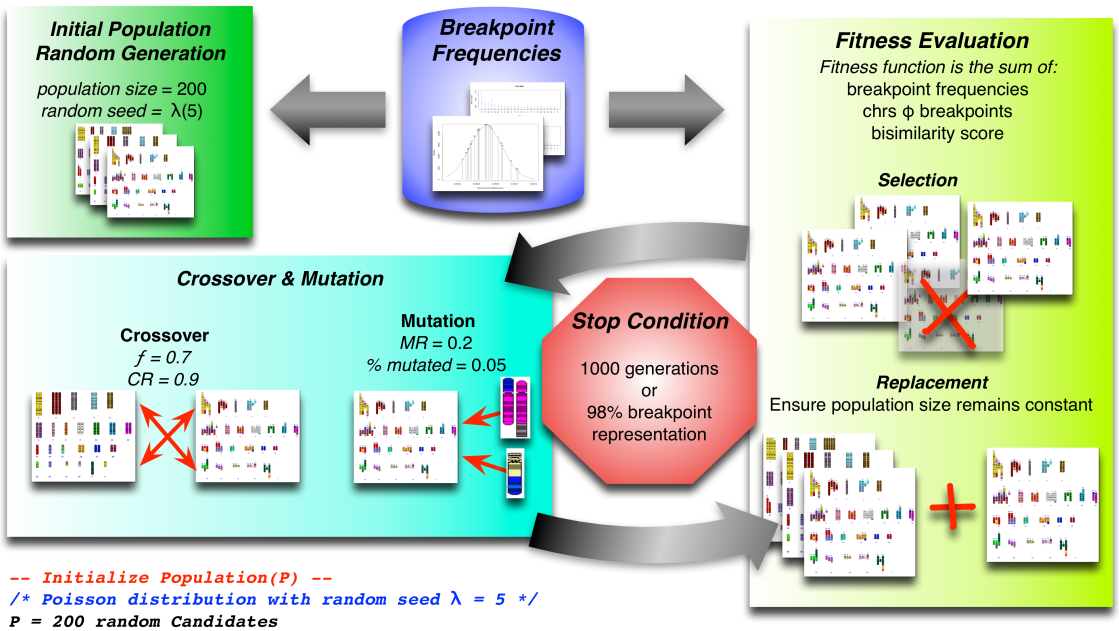
optimal conditions) it would require days of compute time and unrealistic amount of disk space as described above.

Dealing with these limitations required an optimization procedure that selected breakpoints to be tested. Using the raw probabilities from the karyotype analysis to generate the most likely combinations would result in identifying reads that belong to well known breakpoints, while missing those that are less well characterized or unreported in the literature. This means that in order to optimize the selection of simulated references, and avoid bias towards the most commonly known breakpoints (e.g. centromeres are the most reported breakpoints in the microscopic methods, or the Philadelphia chromosome in leukemia), a selection algorithm was introduced to generate populations of breakpoints.

Where there is a large possible search space for a given problem, computational brute force methods that attempt to solve the problem by computing all possible solutions first is rarely appropriate, or practical to achieve. This is particularly true where the characteristics of the solution are not known. Evolutionary algorithms (EA) were developed as a class of search heuristics for such optimization problems. EAs use the concepts of biological evolution (including reproduction, selection, and recombination) to calculate solutions from a population. These require the definition of fitness constraints that are approximations used in the evaluation of each solution. While the concepts are rooted in biological evolution, the algorithms have been successfully used in diverse fields from mathematics and economics, to robotics, chemistry, and biology.

Genetic algorithms (GA) are a subset of EA that use the canonical structure (after initially generating a population) of the form: *1. Calculate individual fitness* → 2. *Select individuals* → *3. Mutate or Cross individuals and Replace* → *4. Repeat from Step 1.* These steps provide a generalized view of what a given GA may look like. One of the common issues found in the use of GAs is the risk of the solution space converging to a local minimum due to the requirement that a new solution improve on the fitness of the parent in order to be included. Differential Evolution (DE) (Storn and Price 1997) was proposed as an approach to avoid convergences in the search process. Practically this means that the diversity of the solution space is prioritized in Step 3 of the canonical GA, and that multiple “best” solutions are possible (or multi objective optimization). It does this by using both *Mutation* and *Crossover* operators to increase diversity before selection. These steps require some parameter tuning for the probability of a given individual in the population to undergo mutation or crossover.

Figure 17 Differential Evolution Selection Algorithm




```
-- Initialize Population(P) --
/* Poisson distribution with random seed λ = 5 */
P = 200 random Candidates

/* Termination conditions have not been met */
while ( generations < 1000 OR breakpoint representation < 98% ) {
   -- Perform Crossover F(0.7), CR(0.9) --
   foreach (Candidate)  {
      foreach (Breakpoint) {
         /* 3 candidates are randomly selected for each Breakpoint */
         if (random num <= CR) {
            x1,x2,x3 = candidate[1,2,3] has breakpoint? 1: 0
            /* add the differences with weighting factor F */
             v1 = x3 + F * (x1 - x2)
               if (v1 >= 0) add Breakpoint to Candidate
               else remove Breakpoint from Candidate
          }
       }
   }
   -- Perform Mutation  MR(0.3), PP(0.05) --
   foreach (Candidate) {
        /* Randomly select P*PP Candidates to mutate */
        if (random num <= MR) add random 1..n breakpoints to Candidate
   }
   -- Select Candidates --
   foreach (Candidate in P) {
      /* Due to the weighting factor of the chr/bp ratio, a low value is higher fitness */
       if (Candidate fitness < max fitness) select Candidate
       /* NCD determines pairwise similarity, used to ensure population diversity */
       if (CandiateNCD >= MaxNCD) select random Candidate from any similar pair
   }
   -- Add new Candidates to Population --
   while (P < 200) add random new Candidate to P
   /* Count number of generations */
   generations += 1
}
```

The selection algorithm is an implementation of differential evolution as this variant of genetic algorithms provides multiple solutions across the search space. The process of DE can be summarized as: 1) generate initial population 2) cross each breakpoint pair by exchanging partners given a crossover constant (CR) 3) mutate each breakpoint pair given a mutation constant (F) 4) evaluate the individual fitness 4) evaluate the population diversity. When either the population diversity reaches a reasonable optimum or a certain number of generations have been run the selection algorithm stops. The parameters CR, F, and maximum generations were all selected to optimize the diversity of the end population. Each of these constants can also have a large impact on the computational time it takes to generate a population.

In the context of the genome simulation for MultiSieve the goal was not to find the most likely single candidate, there is not enough information on structural variations for that currently, instead the goal was to find a diverse population of possible candidates. The information to both create and assess the candidates was provided by the karyotype analysis. However, it was also important that the solution set was not overly biased to only the most probable variants as would be likely if a simple search was performed based on frequencies alone. This means each individual karyotype was assessed in two parts: how similar is this karyotype to all others in the population; and how probable is the individual karyotype.

To answer the "*h*ow similar is this karyotype to all others in the populations" question, the similarity of any single individual to any other was assessed through a weighted graph. Each individual was connected to all other individuals, the edge between each pair was weighted by the Normalized Compression Distance (Bennett et al. 1998; Cilibrasi and Vitányi 2005) or NCD, between each pair of individuals (see Equation 1). The NCD was calculated based on the compressed string representation of the breakpoints (the strings are preformatted by ordering and removing extraneous characters "9p24,4q28"),

Equation 1 Normalized Compression Distance

$$\boldsymbol{NCD = \frac{C_{xy} - min\{C_x, C_y\}}{max\{C_x, C_y\}}}$$

where C_x and C_y are the compression scores of the respective individuals and C_{xy} is the compression score of both together. The closer to zero the NCD is, the more similar the two individuals are.

To address the question "*how* probable is the individual karyotype" the fitness for an individual karyotype was assessed in three parts:

1. Sum of the probabilities for each breakpoint in the individual karyotype. Here a sum that was too high indicated the inclusion of multiple breakpoints with high probabilities (e.g. Philadelphia breakpoints). A high sum for the probabilities is penalized, and was therefore considered to be "poor" fitness.
2. The pairwise sum of NCD scores between the individual and all others in the population. Here again a high score indicated significant similarities between this individual and others and was penalized.
3. The ratio of the total count of chromosomes to breakpoints in the individual, this measure helped to increase the general diversity of the population overall as individual karyotypes that had a 1:1 ratio of chromosomes to breakpoints were penalized (the populations rapidly resolve to every individual having the maximum possible breakpoints without this).

These functions were used in the optimization of the population being searched, in the DE this is the *Selection* step (see Figure 17) where the population is evaluated and filtered for individuals to be used in the next iteration. In the *Selection* step there were two tests applied to the individual solutions. Each individual karyotype was checked for its similarity (NCD) against all other individuals. If one was found to have a NCD below a minimum threshold (0.2 was selected as the threshold that maximized diversity) the pair was run through a Tournament-like selection. Essentially, one of the two was randomly selected to stay in the population and the other was removed in the following generation. The second step was to remove individuals with either "perfect" fitness (e.g. no breakpoints at all) or a score that was too far above a maximum threshold (e.g. meaning that a very high number of breakpoints or chromosomes had been represented in this individual or it was too similar to others in the population).

The algorithm was terminated when the population reached either a maximum number of unrepresented breakpoints across all possible bands (e.g. 3-10), or 1000 generations had been run. With a population of 200 the algorithm generally reached an optimal point with regards to breakpoint representation within 400 generations.

The result is a genetic algorithm with an optimization function provided for the entire population being iterated over, instead of a single solution. This function combined the fitness of all individual references, and a measure of the uniqueness or diversity of the DE population. The diversity score ensured that cytogenetic bands with a smaller probability of being involved in a recombination event could be represented, enabled the testing of chromosomal regions that may otherwise have been underrepresented due to a bias in the frequency data (e.g. missing data for disease-specific aberrations), and avoided over-testing breakpoints that may have been overrepresented in the knowledgebase (e.g. centromeres, Philadelphia chromosome, etc.).

The final output of the selection algorithm was a set of pairs of chromosomal locations that were used to generate a set of miniature references that modeled structural variations. Each of these represented the sequence of the selected recombination. For example t(16;8)(q13;q24), is defined as starting with 16q13 (56700001- 57400000) and ending with 8q24 (117700001- 46364022) creating a recombination point at 700kb from the start of the reference sequence.

The DE selection process was used by MultiSieve to select a diverse (representative) sample of structural variations from an underlying population of known breakpoints. These known breakpoints were retrieved from the "karyotype level" database discussed in section 3.1.1. The next section describes the methods developed to use these mini references to detect actual structural variations in patient samples.

3.2 Alignment, Detection & Scoring of Structural Variations

This section discusses how the large-scale structural variations were identified using the data and selection algorithm discussed in 3.1. This process had three steps:

- *Alignment*: how the bank of small references were used to align the tumor data
- *Detection*: how SVs were detected from these alignments
- *Scoring*: how the detected SVs were scored to identify the most likely candidates.

3.2.1 Alignment

The sequences that were generated, either through random selection or using the DE algorithm, supplemented the standard whole-genome reference alignment. These now functioned as both references for alignment, and models of large-scale structural variation, enabling both an increase in the number of possible alignments and in the computational speed of alignment for testing multiple regions. This meant that MultiSieve was able to quickly search for structural variations that other (e.g. Breakdancer) SV detection tools would be unable to find. This lack is not due to the tools themselves, but rather due to the fact that the alignment algorithms cannot exhaustively report all possible alignments for non-unique reads, and therefore the detection tools have fewer aligned read positions to base identification on.

To align the patient data against the mini references, the reads that were filtered from the original patient as being 'unmapped' (one or both reads in a pair failing to align to a location in the reference genome) or 'discordant' (aligning to two different chromosomes) were aligned to each of these breakpoint model regions using the BWA (Li and Durbin 2009) alignment tool. It is important to keep in mind that by aligning the same reads to multiple references BWA can no longer select a single alignment where multiple alignments are possible. With alignment tools like BWA the reporting of a single alignment for a read-pair is done by either taking the highest map quality score, presuming that the best mapping is correct keeping in mind that this is dependent on the number of mismatches allowed, or by randomly selecting one mapping from the possible alignments. By using multiple references and previously unmapped reads the alignment criteria for greater mismatches and multiple alignments was relaxed. This was necessary due to the potential heterogeneity of tumor samples and low frequency of reads supporting a breakpoint.

3.2.2 Detection

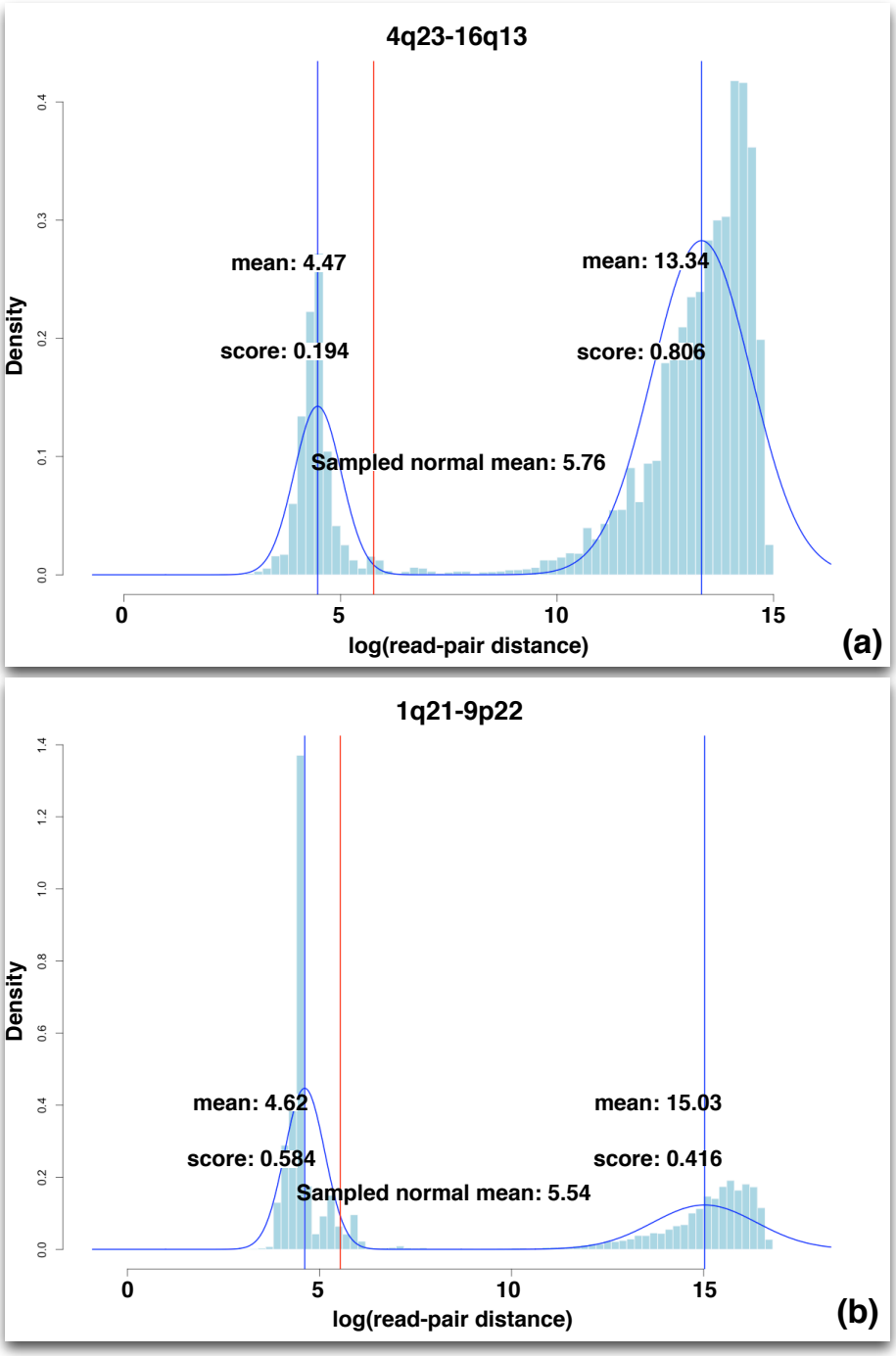
Once the alignment was completed the resulting information was used to detect SVs. Detecting large-scale structural variation using the model regions requires data filtering to remove candidate reads with poor alignments or likely erroneous mappings.

MultiSieve used all discordant and unmapped reads, as there are multiple reasons that a read may have been unmapped in the original alignment step including: higher numbers of mismatches to the reference than were allowed in the original alignment; discordance resulting in one read of a pair failing to align within the time provided by the algorithm; pairs that result from sequencing of a poorly characterized region (e.g. centromeres and telomeres) or other highly repetitive regions (Schbath et al. 2012). This means that when these reads were reported as having aligned to a reference that models a breakpoint region it had to be evaluated and potentially filtered out. Aligned reads for each breakpoint reference were filtered to remove reads where: 50% or fewer of the nucleotides matched based on the CIGAR string (Li et al. 2009); or the summed Phred score was below the mean identified in the original alignment. The remaining aligned reads were then analyzed for the distribution of insert lengths between read pairs.

A non-symmetric bimodal distribution was observed in all read alignments to the in silico references with respect to read-pair insert length. These alignments were investigated using a distribution-based clustering method, expectation-maximization (EM), to cluster reads belonging to each distribution, and to identify the means of each (see Figure 18).

The first distribution was characterized by reads with an insert size near or below 2 s.d. of the mean insert size (as determined from a sampling of aligned reads in the original BAM). Additionally these had a map quality score of less than 30 (Li, Ruan, and Durbin 2008), indicating a higher rate of mismatches or a poorer quality read. It is also important to note that these read-pairs were mapping within the normal range based on insert length. As these read-pairs mapped within the normal insert distance they should have been aligned by the original alignment to the reference, unless these read-pairs happened to align to either side of the breakpoint modeled in the mini reference. However, the synthetic reference models had fusion points at the boundaries of cytogenetic bands. While it is not impossible for a breakpoint to actually occur at this point, it is unlikely to have done so in every reference that is tested. Sampling reads in the first distribution across several of the modeled breakpoint references showed that it was not the case that these “normal” aligned reads were aligning across the cytogenetic band boundaries at all. Thus the first distribution is primarily due to read pair alignments with a high likelihood of error and could not be used for direct identification of structural variation.

Figure 18 Bimodal Distributions in Model Alignment



Distribution of logged read-pair insert distances from aligned reads showing a bimodal distribution with two clear centers. The distribution on the left side is from read-pairs with a "normal" distance (about 400bp for Illumina), the distribution on the right are from read-pairs which span both bands. In (a) the reference is sequences from 5q13 and 8q24 and the aligned reads show a very clear normal distribution centered around a mean of 16 with a large number of supporting reads, suggesting that this is a likely SV. In (b) the second distribution is poorly defined. The reads in this simulated region are less likely to indicate a SV and instead provide a baseline for erroneous discordant alignments. Image published in (Killcoyne and del Sol 2015)

The second distribution consisted of reads that aligned with an insert size greater than 4 s.d. from the mean insert size (Ruffalo, LaFramboise, and Koyutürk 2011). This is a commonly recognized definition for one type of 'discordance' resulting from structural variation. Due to the discordance of the reads in the second distribution there was no calculated map quality score. This made it necessary to define the quality of the read and the alignment by using higher (>50%) CIGAR and summed Phred values (> mean Phred). Therefore this distribution was viewed as representing the possible alignment of read pairs that resulted from sequencing a breakpoint region.

The composition of these two distributions was used to indicate that a breakpoint had occurred, and was used as the basis of the scoring function discussed below

3.2.3 Scoring

Using EM to identify the distributions enabled use of the mixture model parameters to calculate the probability of an aligned read pair belonging to the second distribution (using the R package 'mclust' (Chris Fraley 2002)):

Equation 2: Expectation Maximization Conditional Probability Ratio (EMr)

$$\boldsymbol{EMr} = \frac{\sum_{n=1}^{N} \boldsymbol{P}(\boldsymbol{n}|\,\boldsymbol{z})}{\boldsymbol{N}}$$

where *P(n|z)* is the conditional probability of the nth read belonging to each of the two distributions identified. $\boldsymbol{EMr}$ (see Equation 2) reflected the proportion of reads that were found in the second distribution to have a 'discordant' insert distance. It was derived by finding the probability of the *nth* read belonging to the second distribution, then iterating over the set of *N* where *N* is the total number of reads aligned to this reference (across both distributions). The resulting value was a ratio based on the number of distributions found and the sum of the $\boldsymbol{EMr}$ for each is equal to 1.

As the first distribution described erroneous alignments it was used to find the first cutoff value for further analysis of the alignments described by the given in silico references. All models where the second distribution had an $\boldsymbol{EMr}$ below this cutoff were discarded. It has to be noted however, that as the $\boldsymbol{EMr}$ is a ratio it could not be used without additional information to indicate a structural variation, as a ratio of 0.6 is of less significance if only 100 reads aligned in the second distribution than if 10,000 reads aligned.

Therefore an additional step was included to evaluate the alignments. The second step was to identify putative breakpoint locations based on alignment positions. A sliding-window clustering approach was used in each model breakpoint alignment to identify

regions with an over-representation of reads aligning to a location. The aligned reads from the second distribution that passed the quality filtering steps were clustered by position if the read pairs also spanned both chromosomes represented by the simulated reference. This is an important qualifier. Since the breakpoint references used several kilobases of sequence on either side of the breakpoint boundary it was possible for a discordant read pair to align both reads in one chromosome with a discordant insert distance, which is not what the method was currently trying to identify.

The windowed clustering strategy is a common method for identifying structural variation as well as for providing an estimation for depth-of-coverage (Medvedev, Stanciu, and Brudno 2009). Various methods including BreakDancer (K. Chen et al. 2009), MoDIL (Lee et al. 2009), VariationHunter (Hormozdiari et al. 2010), and Pindel (Ye et al. 2009) use clustering or window-based strategies to detect signatures for structural variants. Exactly what signatures they search for and the method used to cluster the reads differs based on the variants (e.g. Pindel and MoDIL detect small indels, BreakDancer and VariationHunter may detect multiple variants including inversions, deletions, and translocations).

In MultiSieve the positional window clustering enabled each location to be assessed for coverage depth. However, the distribution of this coverage was also important. If the aligned read locations were uniformly distributed across the breakpoint reference all of the positions would represent poor alignment sensitivity. This means that it was the outliers that are likely to be representative of actual alignment across a breakpoint. In order to adjust the ratio derived from EM, the major outlier in the positional window-clustering was used to determine the proportion of reads in the second distribution which potentially indicated large-scale structural variation (see Equation 3):

Equation 3: Combined distribution and positional clustering for translocation scoring (Tx)

$$Tx = \sum \left(EMr, \frac{W_{max}}{N_b} \right)$$

where W_{max} is the cluster with the highest total count of reads from the second distribution, and $N_b$the total number of aligned reads within the second distribution which passed quality filters. The maximum cluster proportion was used by MultiSieve as the feature most descriptive of structural variation based on simulations in real patient data.

The main approach taken when researching the methods that went into the MultiSieve method was to develop a SV detection approach that can quickly search cancer samples and find structural variations that other tools were not able to find. The methods used to do this combine modern HPC approaches to scale out the problem, the use of background information and optimization strategies to limit the search space, and the

actual implementation of a system to discover and inform about detected SVs. Validation of the method using synthetic data with known characteristics is discussed, alongside real world (e.g. patient data) examples, in the Results chapter.

3.3 Distributed Computing for Genomics

Handling genomic sequence information requires that the scale of the data is a major consideration in any type of analysis. A single BAM file for a human whole-genome sequence is billions of reads and hundreds of gigabytes. Even the reference genome, which is simply 3 billion base pairs or characters in a file, is several GB in size. Altering that reference for both small mutations (SNVs, indels) and structural variations (inversions, translocations) results in millions of computations along with the necessary computational work involved in reading and writing the results of these computations to a new reference file. Doing this efficiently and rapidly requires the use of high performance computing, typically through the use of distributed frameworks.

The applicability of these highly distributed computing environments was tested to solve two specific issues: how to generate highly mutated reference sequences (both for bench marking and for the MultiSieve reference alignment); and how to store and retrieve large numbers of reads, which is needed to enable scaling of MultiSieve methods to the level of populations of patients. These scalability issues needed to be addressed if the in silico reference solution discussed in this thesis was going to be able to be used on populations of genomes.

The most suitable distributed framework for these specific problems was identified, and the corresponding analysis system was developed. For the mutation of large numbers of references MapReduce was used (see Section 3.3.1), and for the population scale genome read data warehouse (see Section 3.3.2) MongoDB was the most appropriate. The Amazon Elastic Cloud Compute (EC2) platform and web services provided the means to test the scalability (scale out) of these systems.

3.3.1 HPC for Mutation using MapReduce

The method discussed in this thesis required the generation of a large number of sequences, which are mutated versions of the reference sequence. The MapReduce framework was used to develop both in silico reference sequences (Killcoyne and del Sol 2015), and also to develop novel mutated references, which were used for benchmarking and evaluation. When developing the mutated reference sequences, a number of steps were undertaken for each kilobase of sequence as part of a pipeline called FIGG (Killcoyne and del Sol 2014): determine the GC content of each segment; apply the pre-calculated mutation rate for each mutation class (e.g. SNV, indel, substitution); generate a

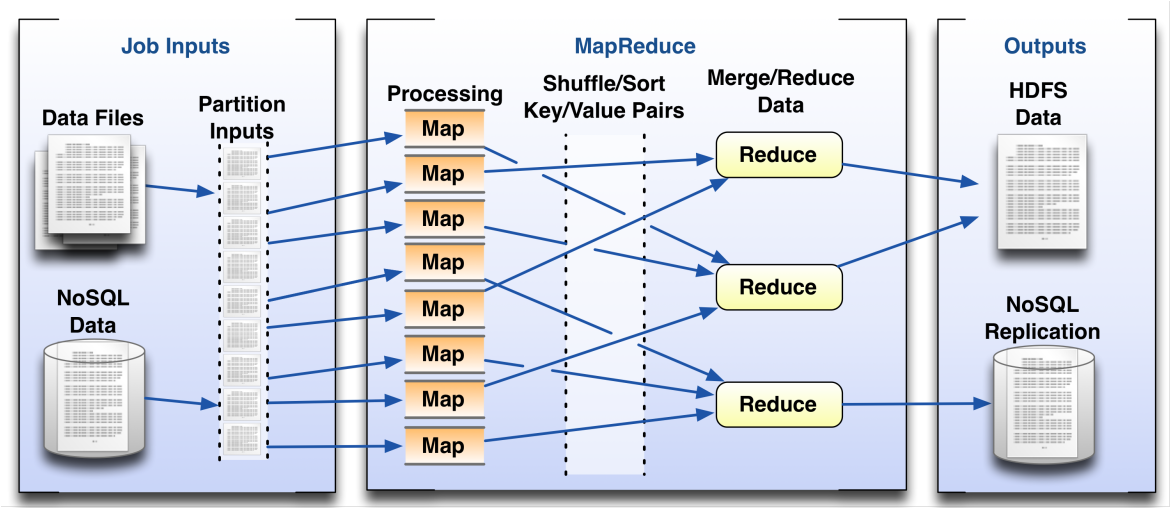
new sequence with the synthetically created mutations. If large-scale variations were also being applied, sequence from other chromosomes may be inserted or large sections of the sequence may be reversed. Sequentially mutating a genome from chromosome 1 to Y requires $3x10^6$ independent processing steps. In order to perform this task the computational load was distributed through the use of high performance computing (HPC) frameworks and compute clusters.

HPC enables the rapid calculations necessary for mutating reference sequences through the aggregation of computing power as well as the distribution of memory, storage, and computation. While there are various solutions available for parallelizing computation MapReduce (Dean and Ghemawat 2008), first developed by Google to handle large unstructured data, has been adopted in bioinformatics methods more than any other solution. This is largely due to the simplicity of development as well as its rapid acceptance as the standard “cloud computing” platform by major cloud vendors like Amazon, Google, and Microsoft. To generate the in silico references Hadoop MapReduce on Amazon EC2 was used, as this provided the most practical and scalable infrastructure.

The major advantage of MapReduce is its simplicity, as it enables tools to be developed quickly and scale-out effectively. Unlike many other parallel frameworks that require developers to manage communications within the processes, MapReduce (see Figure 19) breaks down the computation into two major phases, *Map* and *Reduce* that are automatically executed in parallel across the available compute resources (nodes). This separation between the processes enables the developer to (nearly) exclusively concentrate on the process or analysis.

The *Map* phase takes the input data that is pre-partitioned by the framework using a provided data reader (e.g. tab-delimited text, HBase, per-line text) that is customizable if required. Each *mapper* runs the computation provided by the developer over its chunk of input data. The output is a key-value pair defined by the application that defines some relationship between the data. For example, if the computation was to count all occurrences of G and C nucleotides in a sequence string it could output the key-value (GC, 12). Or it could output more complex information as an object as in the case of a sequence mutation where the key is the location and the value is an object that includes a sequence, location, and GC content. After the mappers have completed their tasks there is an intermediate *shuffle/sort* phase to sort all of the outputs with the same key into a single group. Each group of keys is then provided to a *reducer* that can do further computation over the grouped values (e.g. summing all of the GC values) and outputting it to a file or database.

Figure 19 MapReduce



Schematic overview of MapReduce. Input data is partitioned into chunks automatically and sent to mappers. Each mapper performs the computation and outputs a key-value pair that are sorted and send to reducers. Reducers can apply some final computation and output the data to a file on HDFS or a NoSQL database.

Apache Hadoop and HDFS (Highly Distributed File System) provides an open-source version of MapReduce that can be used on internal clusters, or on the Amazon EC2 platform and are widely used within bioinformatics.

Since the Human Genome Project published the draft human genome sequence the cost for a single genome has continually decreased, while the technologies used have continued to improve. The result is that the pace of improvement in sequencing as well as the number of genomes being sequenced has hugely outpaced the ability of a single computer to align and analyze them. This has led to the use of distributed processing architectures across a range of genome alignment in analysis applications (Schatz, Langmead, and Salzberg 2010). Apache Hadoop (The Apache Software Foundation) and MapReduce has been the most commonly used framework for scaling up genome analysis via HPC. This is due in large part to its relatively simple workflow. Unlike earlier parallel computing frameworks the software development required is relative minor. In many cases scripts that manage the input and output data (e.g. sequence reads and aligned sequences) may be used via the Hadoop "streaming" utility, which manages all of the distribution of the data and computing nodes, using the analysis script as a mapper or reducer as indicated by the developer.

Hadoop MapReduce has been successfully applied to a number of genomics analysis projects. In 2009 the first projects to make use of these frameworks were released. Cloudburst (Schatz 2009) provided a read alignment implementation on Hadoop MapReduce which improved on the sensitivity of alignments by allowing more mismatches

or reporting every possible match. Crossbow (Langmead et al. 2009) used the framework in SNP detection, this time on Amazon EC2 which enables the rapidly scaling of hardware resources. In 2011 SAMQA (Robinson et al. 2011) was released, building again on the Hadoop MapReduce platform to rapidly assess the quality of high-throughput sequencing data post-alignment, and in 2012 the Hadoop-bam tool (Niemenmaa et al. 2012) was released, building the popular Java implementation of SAMtools (Picard) (Li et al. 2009) into the MapReduce framework. Many other genomics tools have since been developed or adapted into MapReduce for rapid processing of sequence data.

In the context of the genome mutation simulation software (FIGG) it made sense to distribute the processing and the computational work involved in mutating the sequence could be easily mapped into the *Map – Shuffle/Sort – Reduce* pipeline. In the *Map* step the genome is broken into multiple chunks, which are mutated in parallel. In the *Sort/Reduce* step the mutated sequences are resorted into chromosome order, including any large-scale structural alterations such as kilobase size inversions, insertions or translocations and finally output to a FASTA or the Hadoop distributed database, HBase. This process scaled nearly linearly with the number of cores made available to Hadoop, and so provided the many hundreds of mutated references required.

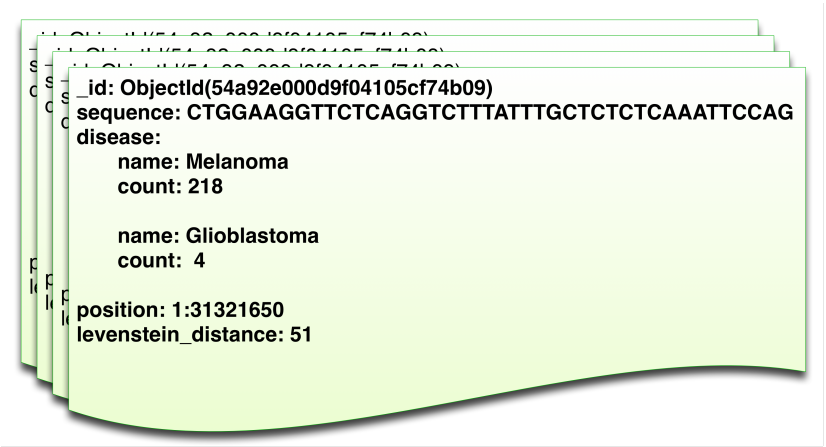
3.3.2 HPC FOR READ WAREHOUSING & SEARCH USING MONGODB

One of the issues related to variant identification in large-scale data is the searching through billions of read pairs with their associated positional information. This search is neither an easy or quick process. Most of the current variant identification algorithms for both small and large variations take many hours to analyze a single whole genome sample. Comparing across multiple samples requires that each sample be integrated individually into an analysis. This means adding a new genome, a new variant, or finding a variant that may have been previously identified can involve reanalyzing the entire set of genomes each time. It also means that data about the read itself is lost as the analysis must compress the aligned reads and report only the variation from the reference. Considering the low frequency of reads supporting structural variations in particular, being able to analyze across many patients simultaneously could help to identify driver events in the structural changes to the genome. In order to do this all of the reads for every patient needed to be available and easily searchable.

Searching across hundreds or thousands of patients, each with billions of reads per genome, also required the use of highly distributed computational solutions. A data warehousing strategy using a NoSQL *document based* distributed database (described in the Introduction chapter, section 1.4.1), MongoDB, was prototyped to enable this. This data warehouse scaled directly by simply adding more nodes to the system using a

sharded cluster. Sharding breaks up the data so that each shard stores only a part of it (e.g. shard-1 may store records from A-C, shard-2 from D-F, etc) and then queries that match a specific shard only need to be run on that shard, rather than the entire dataset.

Figure 20 MongoDB sequence read document



The design of this document is specifically aimed at querying reads by disease type, disease frequency, sequence motifs, or simply chromosome location. As each field can be indexed it is simple to search across this data without first normalizing across multiple relational tables.

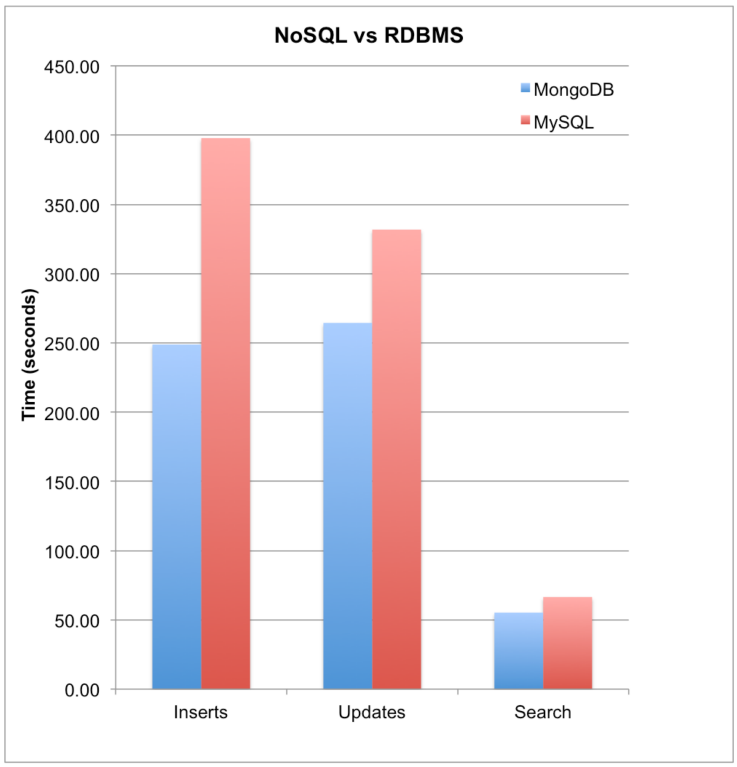
The initial document design (see Figure 20) was planned to test that the system could scale effectively to address population scale analysis. In this way the system had to offer scalability for multiple samples of sequence data. Additionally the system was designed to ensure that the read data could be de-identified from the specific patient for genomic data security concerns. Using a document-based solution had a number of distinct technical advantages:

- It provided the ability to index and query on any field or subfield in the document, which is not easily done in key/value systems where the key is the index, or in graph systems where the fields may become either relationships or nodes.
- Each document could include any number of different fields, which in this case enabled the addition of new metadata to reads, such as clinical significance.
- It would be relatively simple to add other types of motif information for transcription factor binding sites, methylation sites, tissues, or cell types without requiring the addition of new tables that complicate the queries and increase the size of the database itself.

To be useful the test system needed to be able to actually search the sequence reads directly. To achieve this a specialized search algorithm was developed.

The sequence search database was sharded by the first nucleotide into A, C, T, or G shards. Each read that was to be loaded was compared to an initial (or “root”) read and the Levenshtein distance (LD) (Levenshtein 1966; Damerau 1964) between them was calculated. The read was then added to the correct shard and the distance is included in the document.

Figure 21 NoSQL vs RDBMS



The NoSQL solution (MongoDB, blue bars) was faster when compared to a sharded and clustered version of the relational database MySQL (red bars). This is due to the fact that each process required multiple joins across relational tables in MySQL while MongoDB only needs to interact with a single document at a time.

Once the load procedure had been optimized to effectively distribute the data across the shards, a search algorithm was developed to use this information to enable effective search. The search algorithm used four steps:

1. Select the appropriate shard based on the first nucleotide of the query read: A,C,T,G
2. Calculate the LD between the “root” (e.g. the first sequences loaded) and the query read
3. Select the group of reads from the database with the same LD as the query read
4. Compare the query read to the reads with the same LD distance by directly comparing character-by-character and stopping the instant a mismatch is detected.

This search approach was tested against a commonly used string matching algorithm (Aho and Corasick 1975) and was found to be consistently faster as the data sizes increased from 600MB to 1GB BAM files.

Using the search approach a sharded MySQL relational database design using three relational tables was also tested against the MongoDB design in a cluster on Amazon EC2. The distributed database MongoDB was consistently faster in each test (see Figure 21): initial inserts of 700K reads were 37% faster than MySQL; record updates were 20% faster; and a search without altering or adding any records was 17% faster.

It would be expected that the relational database solutions, even with sharding, would be slower as the scale of data increased, due to the fact that each process required joining three separate tables. Therefore this supported the need to use distributed data warehousing solutions in order to enable the types of queries and the addition of new metadata described above.

3.4 Chapter Summary

This chapter described the data, HPC frameworks, and analytical methods that were used in the development of the MultiSieve method. The primary aim of which was to enable detection of structurally variant regions in cancer genomes by using multiple references that model probable SVs to realign the discordant or poorly reads from aligned cancer sequences. To generate these sequences a knowledgebase of large-scale breakpoint frequencies in cancer is used to inform a search optimization algorithm that first selects, then outputs new “mini” references that include a selected breakpoint.

MultiSieve then aligns and scores each reference alignment using a score ($\boldsymbol{Tx}$) that is based on the distribution of reads that have aligned to the reference, and a windowed positional clustering of the aligned reads.

In the Results chapter the analysis of karyotypes to determine breakpoint frequencies for the optimization algorithm is described. A validation test is performed and the false positive rate for selection of the best $\boldsymbol{Tx}$ scores (and therefore structurally variant regions) is selected. This is then applied to the analysis of nine patients across seven cancer types from TCGA patient data, and compared to a common SV detection algorithm. Finally, the HPC methods described here were tested for scalability using the Amazon EC2 services with whole-genome data that simulated actual mutation frequencies.

CHAPTER 4 RESULTS

This chapter describes the results of each analysis leading up to the identification of structural variation in patients. This analysis builds upon the previous work described in the Methods chapter.

The Breakpoint Analysis section (4.1) details the aberrations that are defined in the publicly available karyotype data. It further describes the analysis of chromosomal instability that was performed at both a genome-wide and per chromosome scale, as well as the final output data that is used in the generation of *de novo* references. In the Tx Score Validation section (4.2) an analysis of simulated translocations is performed, and estimated sensitivity is reported. These results provide the basis for the subsequent section, Selecting Appropriate References (4.3) where the differential evolution algorithm is used to optimize the search space. Finally, these references were used to analyze TCGA patient data to find regions containing large-scale structural variation in Patient Data: Germline/Tumor pairs (4.4) and compared to the most commonly used SV detection tool.

The final section, Population Scale Analysis (4.5) outlines the analysis performed on using suitable HPC frameworks to ensure that the system can scale to the level of populations of genomes. This work includes details of small variant analysis for both normal and cancer patients, and the corresponding suitability of the MapReduce method for generating whole-genome simulated samples is shown.

4.1 BREAKPOINT ANALYSIS

An analysis of breakpoints reported in publicly available karyotype data was undertaken to provide a knowledgebase from which to generate simulated reference models of structural variations (as described in Methods section 3.1.2). It was presumed that not all regions of the genome are equally likely to break as fragile regions and microtubule defects are known to influence chromothripsis. This analysis was approached in three parts: a general frequency analysis of karyotype aberrations; structural instability; and the influences of specific regions on general stability.

4.1.1 ABERRATION FREQUENCY ANALYSIS

The karyotypes used in the breakpoint analysis were primarily collected from patients and curated by the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (e.g. 99% of the karyotypes). The remaining karyotypes were cell lines and some patients found in the NCBI SKY/FISH database. Identification of breakpoints was performed after the parsing and cleaning step described in the Methods chapter, section

3.1.1. The definition of a breakpoint in a karyotype was provided by an aberration that was described with a chromosome and cytogenetic band (see Table 4). The remaining aberrations described in a karyotype were amplifications and deletions of chromosomal material, either through aneuploidy or extrachromosomal DNA from ring or double minute chromosomes.

Table 4 Example Aberrations

ISCN ABERRATION	DESCRIPTION
t(9;22)(q34;q11)	Translocation between chromosomes 9 and 22 at band 9q34 and band 22q11. Also known as the "Philadelphia chromosome"
inv(12)(p13q15)	Inversion of the segment from p13 to q15 in chromosome 12
del(14)(q21q24)	Deletion of the segment from q21 to q24 in chromosome 14

Examples of aberrations as described by the ISCN standards and a description of the defined aberration.

To recap, from the Karyotype and Mutation Analysis (section 3.1.1 in Methods) there were 30,558 unique aberrations that defined unambiguous breakpoints. Each of these was classified by the type of aberration it represented (see Table 5), 83% involved a translocation making it the most commonly identified aberration in the karyotypes.

A frequency analysis of the breakpoints identified in the aberrations described above found all cytogenetic bands involved in one or more aberrations. The frequency was normally distributed and every band was found in multiple aberrations ranging from a minimum of 15, to a maximum of 682. With a frequency range this large from data with a huge quality variation in regards to both the reporting and the curation it was necessary to find out what this meant at both a genome-wide and a chromosome specific level. At this data scale it was possible to identify a measure of chromosome instability based on general and structural information.

Various measures of instability have been developed based on small variations such as SNPs which result in loss of heterozygosity (LOH), or in short repeats resulting in MSI. Chromosomal instability can involve polyploidy through the gain or loss of whole chromosomes due to a segregation error, or in the duplication, deletion, or translocation of segments of a chromosome. Each of these will result in a breakpoint at the terminal regions of the specific aberration, and if an insertion or translocation occurs, at the point where the sequence is added. Analyzing these regions to identify patterns of whole chromosome instability based on karyotypes provided a measure of large-scale variability.

Table 5 Aberrations Found in Public Karyotypes

TYPE	COUNT	BP	DESCRIPTION
add	756	✔	Added chromosomal material to the telomere regions
del	1927	✔	Deleted chromosomal material
der	12409	✔	Derivative chromosome, typically a novel chromosome involving multiple aberrations
dic	973		Dicentric chromosome, e.g. two centromeres
dmin	5		Double minute chromosome, extrachromosomal DNA in a ring without a centromere
dup	628	✔	Duplicated section of a single chromosome added to the telomeric region
gain	24		Aneuploidy, whole chromosomal duplication
ins	38	✔	Added chromosomal material, typically between two bands within one of the arms
inv	837	✔	Section of the chromosome that has been inverted
iso	116	✔	Isomerization of the chromosome around one centromere of either arm
loss	24		Aneuploidy, whole chromosome loss
ring	149		Chromosome where the arms have fused to create a ring
trans	8605	✔	Inter-chromosomal translocation of material from one chromosome to another
unk	4113		Aberrations that are not defined by the ISCDN standards, or are too incorrect for the parser to determine the aberration

Describes the aberrations discovered in the available karyotypes. The column 'BP' indicates whether the aberration can be analyzed for breakpoints.

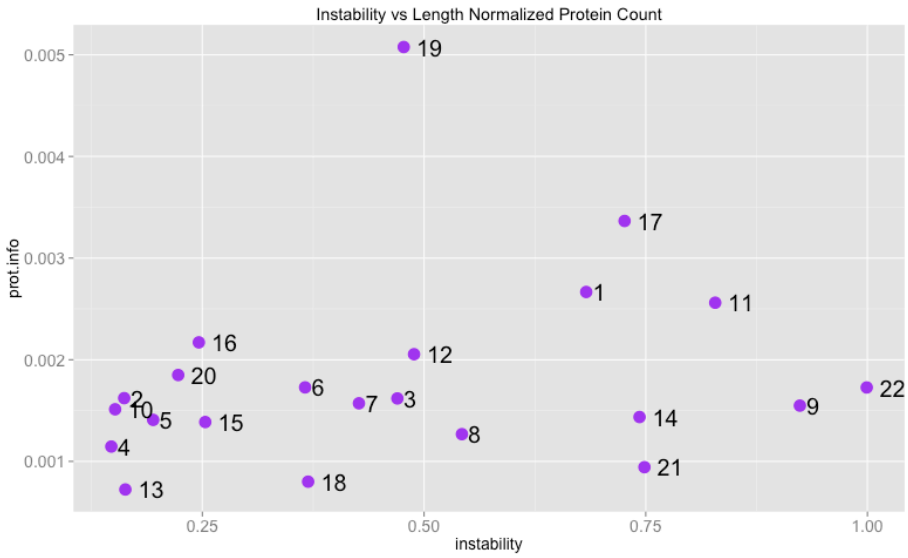
4.1.2 CHROMOSOMAL INSTABILITY

Understanding large-scale instability required comparing breakpoint frequencies, calculated at a cytogenetic band scale (e.g. many kb to Mb in length) with chromosomal data at a similar scale. Ensembl statistics at the band and chromosome level regarding base pair length, protein-coding genes, numbers of small variations, and RNA coding genes (build GRCh37 analyzed August 2013) were used to calculate chromosome level instability as a function of the information content of the chromosome or band (e.g. number of genes).

The breakpoint frequency per chromosome is strongly correlated with the base pair length of the chromosome (r=0.46, p=0.03). This was expected as the breakpoints are demarcated in karyotypes by the cytogenetic bands and the longer a chromosome is the more bands it exhibits. There is also a strong correlation with the number of genes (*r=0.59, p=0.002)*, which is related to the base-pair length in most chromosomes. The

exception to this is chromosome 19, which has the highest gene density in the genome (30.5 genes per Mb). In order to correct the length bias in the frequencies, the length of each chromosome is adjusted non-linearly and the frequencies are calculated using the adjusted length.

Figure 22 Chromosomal Instability Scatterplot



Shows a scatterplot of instability scores vs length adjusted protein coding genes per chromosome. Chromosome 19 is an outlier due to its gene density, which is not fully corrected using the non-linear adjustment. Chromosomes 9 and 22 are at the highest end of the instability score, driven primarily by the Philadelphia breakpoints 9q24 and 22q11.

The adjusted frequencies were used to calculate instability relative to the information content (e.g. protein coding genes) of the chromosome. As these frequencies were normally distributed an "instability score" was calculated based on the probabilities of the scores occurring within the distribution (see Equation 4).

Where I is the instability score defined for each chromosome as the sum of breakpoints $\sum_{bp}\binom{bp}{chr}$ per the adjusted chromosome length $L_{Chr}^{0.7}$, taken as a probability of occurrence within the distribution.

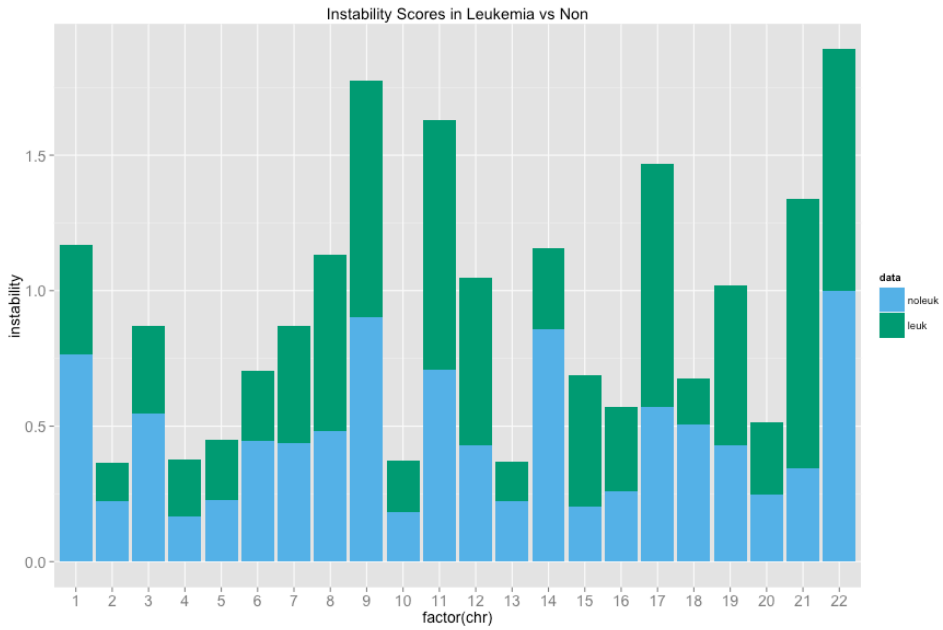
Equation 4 Chromosome Instability Score

$$I_{Chr} = \left\| \frac{\sum_{bp}\binom{bp}{chr}}{L_{Chr}^{0.7}} \right\|$$

In this measure chromosome 19 was an outlier, due to the high gene density, without it the correlation between the instability score and the information content, defined as the length normalized protein count ($L_{Chr}^{0.7}$), improved (from *r=0.23* with chromosome 19 to

r=0.35 without*)*. When the same analysis was performed excluding all karyotypes from leukemia patients (see Figure 22) the most unstable chromosomes were 9 and 22, which are involved in the Philadelphia chromosome and found most often in leukemia. However, the chromosomes that showed the largest change between the full dataset and the non-leukemia dataset were 11, 17, and 21 suggesting that their instability was driven primarily by leukemia (see Figure 23).

Figure 23 Leukemia vs Non-Leukemia Instability Difference

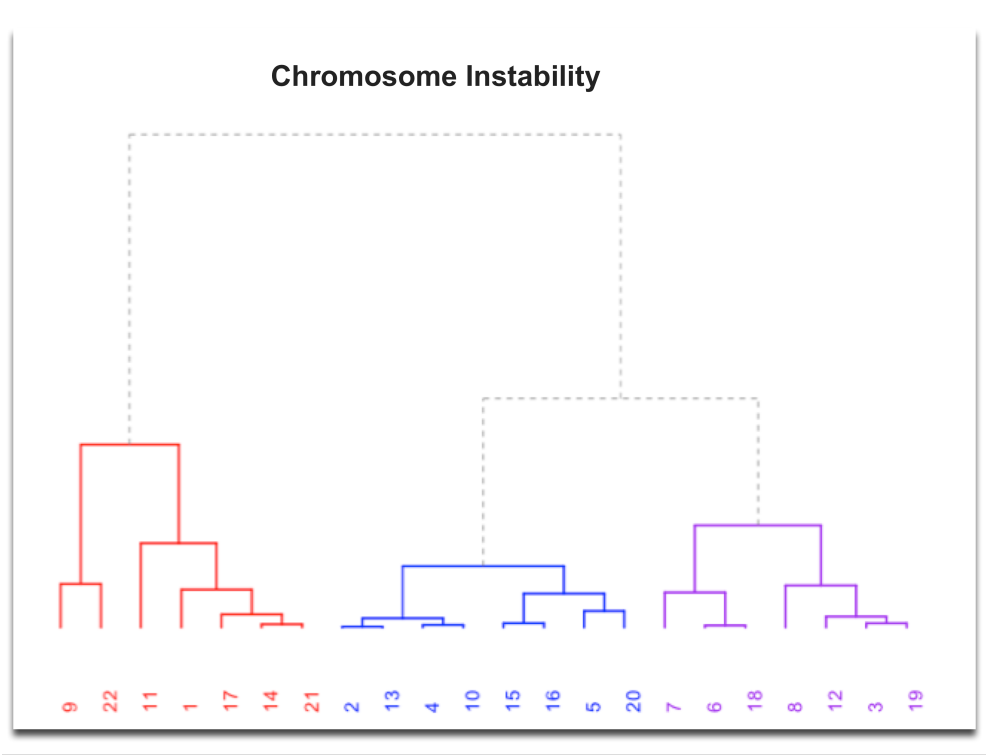


This plot shows the comparative contribution to instability between leukemia (green, top bar) and non-leukemia (blue, bottom bar) samples. In most chromosomes the contribution of the non-leukemia samples is greater. Those chromosomes with aberrations found frequently in leukemia (9,11,17,22) show equal or greater contribution by the leukemia samples. However, sampling the leukemia at rates comparable to the other cancers did not significantly alter this. This could be that the curation of the karyotype data is poor, or that these are relatively common in other cancer types.

Clustering the instability scores (see Figure 24) showed the influence of the leukemia breakpoints on general instability. The first cluster included 9 and 22 (Philadelphia breakpoints) as well as 11, 17 and 21. When the instability scores were calculated and clustered by excluding all leukemia's only two chromosomes were assigned to different clusters (chromosomes 17 and 21). However, the results for the chromosomes that were represented in the Philadelphia translocation (9 and 22) were always the same. Additionally, the Philadelphia breakpoints (22q11 and 9q34) were the most frequently found breakpoints in both the datasets with and without leukemia. It must be noted that while this analysis tried to correct for the specific bias due to cancer type, it could not correct for observational bias that may be part of this data. The Philadelphia chromosome

is perhaps the best-known and oldest example of a structural variation with observable effects in a patient. Therefore it is probable that experimentalists and pathologists were more likely to identify and report it in patients, thus a bias is possible, as the curated data would include these breakpoints more often.

Figure 24 Clustered Instability Scores



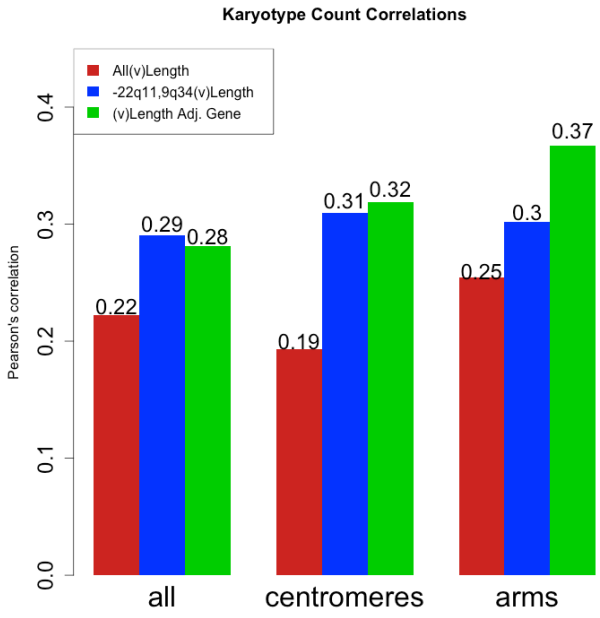
Clustered instability scores using karyotypes from all cancers shows three clusters. Group 1 (red) includes 9,22,11,1,17,14, and 21. All of these are highly represented in leukemia's. Group 2 (blue) includes 2,13,4,10,15,16,5, and 20. Group 3 (purple) includes 7,6,18,8,12,3, and 19. Of the top 10 breakpoints by frequency, 6 were represented in the first group and included the breakpoints involved in the Philadelphia chromosome. X and Y are left out of the clustering, but have instability scores calculated.

What this analysis demonstrated is that there were cancer-specific biases in karyotype level breakpoint identifications. In the dataset available leukemia drove most of the differences. However, the karyotypes represented in these data were curated from literature. The observational biases in literature and pathology, due to the ease of collecting samples from peripheral blood, mean that leukemia's are the most common cancers with reported karyotypes. While this bias could not be fully corrected in this analysis, filtering out karyotypes that were correctly recorded (as described in 4.1.1 Aberration Frequency Analysis) enabled a level of correction.

4.1.3 Regional Influence on Stability

Various explanations for structural instability have been proposed ranging from microtubule defects to hypomethylation, microsatellite repeats, and CpG island methylation (CIMP). These features are useful in the investigation of instability and structural variation in a specific region of the chromosome. For instance, chromosomal numerical instability (CIN) and MSI are known to result in distinct phenotypes in cancer but are rarely found in the same patient (Dunican et al. 2002). A similar situation emerges when investigating CIMP in cancer (Issa 2004), it appears to be inversely related to highly structurally variant phenotypes in colorectal cancers (Cheng et al. 2008). However, these features are found at a different scale from karyotypes (e.g. CIMP is base-pair resolution, while a cytogenetic band is several kb to Mb in length). Other regional effects are comparable at karyotype scale, specifically distance from centromeres and the centromeres themselves.

Figure 25 Karyotype Count Correlations



Correlations for each dataset. The red bars show correlation of the bands vs the length of the bands, blue bars show the correlation with the Philadelphia breakpoints removed as they were outliers in every test, and the green bars show the correlation of the length adjusted counts vs the information content (genes) of each band. The groups are 1) all bands, 2) only centromere bands, and 3) all bands except centromeres. Centromeres clearly influence the instability with regards to the number of genes as most centromeres are gene poor regions.

Generally centromeres are poorly characterized due to their highly repetitive sequence (International Human Genome Sequencing Consortium 2004). Their role in mitosis is central to the correct segregation of chromosomes and they are therefore implicated in whole chromosome aneuploidy (Duijf and Benezra 2013). As aneuploidy is a feature of most tumors, one possible result of this instability could involve breaking the chromosome due to microtubule or centromere defects (Burrack and Berman 2012; Janssen and Medema 2012). If this is the case then centromeres could drive most of the instability found in the karyotype data. This was investigated in the datasets that included leukemia as the primary analysis found little difference in excluding karyotypes.

The correlations described at the whole chromosome level were investigated further, specifically whether the length and information content (genes) of the bands were similarly correlated. Again both the length of the bands and the number of genes in each band were correlated with the number of breakpoints reported in the karyotypes for each band (see Figure 25, 'all'). Adjusting the per-band counts by the non-linear length of each band also removes the correlation with length and still correlates with information content. This provided sufficient evidence that specific regions of the chromosome may influence general chromosomal instability.

Across all of the chromosomes the two structures or regions that are shared at a karyotype level are the centromeres and arms. In a karyotype the centromere is simply defined as the region where the chromosome constricts. Therefore the bands that define the cytogenetic region of the centromere are those nearest the constriction and these have been designated as band '11' on both the q and p arms of the chromosomes. In some chromosomes these flanking bands are relatively long and contain known genes (e.g. Xp11 is 23Mb long and has 187 genes) while in others these bands are nearly fully contained within the constricted region and have few or no genes (e.g. 15p11 is 10Mb and has no genes). These regions are implicated in numerical CIN due to microtubule attachment abnormalities resulting in poor segregation. It is possible that more breakpoints occur in centromeres due to kinetochore/microtubule defects. Comparing the number of breakpoint events in each centromeric band to the length as was done for the whole chromosome analysis above, showed a poor correlation. However one of the Philadelphia breakpoints (22q11) appears to be driving the correlation. Removing it significantly increases the correlation (see Figure 25, 'centromeres').

After the centromeres were removed and the remaining bands were tested (e.g. those exclusively in the *p* and *q* arms) for correlations, a significant correlation between the number of breakpoints found in the karyotypes and the length of each band was still apparent. Here again, one of the Philadelphia breakpoints (9q34) skewed the correlation and removing it increased the correlation slightly. However, the information content in the

arms was what drives most of the instability found in the initial CIN analysis (see Figure 25, 'arms'). This is primarily due to the fact that centromeres are gene poor regions, despite this they showed a high rate of instability overall. In both the centromeres and the arms removing the data that is known to be from leukemia patients did not alter the correlations.

Based on this analysis two sets of probabilities were derived (see Table 6): the probability for a specific band breaking with respect to the entire genome (from the initial whole CIN analysis); and the probability of each band breaking with respect to only the other bands within the chromosome (based on the regional CIN analysis).

Table 6 Example Breakpoint Frequencies

HR	BAND	P(BP)	CHR P(BP)
1	p11	0.0066	0.0895
1	p12	0.0045	0.054
1	p13	0.0054	0.0708
1	p21	0.0027	0.0233
1	p22	0.005	0.0632
1	p31	0.0023	0.0174
1	p32	0.0044	0.0517
1	p33	0.0025	0.0202
1	p34	0.0034	0.0338
1	p35	0.0025	0.0203
1	p36	0.0049	0.0612
1	q11	0.0065	0.0884
1	q12	0.0047	0.0561
1	q21	0.0065	0.0881
1	q22	0.0042	0.0478
1	q23	0.0034	0.0339
1	q24	0.0023	0.0171
1	q25	0.0031	0.0296
1	q31	0.0023	0.018
1	q32	0.0036	0.0371
1	q41	0.0021	0.015
1	q42	0.0028	0.025
1	q43	0.0019	0.0136
1	q44	0.0028	0.0248

Example of the result of the instability analysis for a single chromosome. Probabilities that each band would be involved in a break/recombination event were calculated both cross-genome (P(BP)) and within the chromosome (CHR P(BP)).

The optimization algorithm described in the Methods chapter (section 3.1.2) used these probabilities to select regions for use in *de novo* reference generation.

4.2 Tx Score Validation

The *de novo* references generated from the instability analysis described in previous sections enabled MultiSieve to generate regions that modeled probable SVs. These were used to re-align the unmapped and discordant reads, and Tx scores were derived from the resulting alignments. These Tx scores were used as reliability measures of the likelihood of the SV occurring in the region defined by the *de novo* reference.

Table 7 Simulated Breakpoints for Validation

Pairs	Chromosomal Locations	
10-9	10p14(10570829-10571629)	9q21(71837749-71838549)
11-1	11q23(117616704-117617504)	1p32(59598455-59599255)
11-16	11p14(25584895-25585695)	16q23(82337644-82338444)
14-4	14q31(88831369-88832169)	4q33(171226103-171226903)
14-X	14q13(33913464-33914264)	Xp22(24725149-24725949)
15-6	15q23(68259567-68260367)	6q14(87286282-87287082)
17-2	17q24(70706210-70707010)	2p14(65825885-65826685)
18-10	18q21(46083291-46084091)	10q21(64596245-64597045)
19-2	19q13(53054071-53054871)	2p25(8055945-8056745)
2-3	2q33(207285260-207286060)	3q22(135857871-135858671)
2-4	2p23(26014296-26015096)	4p16(10575640-10576440)
2-5	2q22(147702888-147703688)	5p15(10424844-10425644)
3-13	3p25(16246125-16246925)	13q21(68227569-68228369)
4-X	4q22(90104120-90104920)	Xq21(84389111-84389911)
5-19	5p14(23567847-23568647)	19q13(44423925-44424725)
5-X	5q35(172438308-172439108)	Xp21(26879371-26880171)
8-15	8q21(82604665-82605465)	15q15(40553539-40554339)
8-6	8q13(72263564-72264364)	6q25(153682115-153682915)
8-X	8q23(110732717-110733517)	Xq24(116772533-116773333)
9-4	9p24(5818205-5819005)	4q28(128137247-128138047)

Randomly selected chromosome pairs, and the breakpoint locations within each. Note that only 800bp upstream and downstream are used for the breakpoint. This is due to the limitations of short-read sequencing. As the insert size is 400bp or less reads spanning a breakpoint, or being split by the breakpoint will not be found outside 1000bp. A total of 1600bp ensures that reads will be generated that span the breakpoint, are split by the breakpoint, and may not include the breakpoint at all.

In order to validate the scoring method used to calculate the $\boldsymbol{Tx}$ score, simulated data was generated to model inter-chromosomal translocations. First 20 pairs of chromosomes were randomly selected. Within these pairs one location from each chromosome was also randomly chosen as the breakpoint and a FASTA file was generated for the sequence that would result from a merge point between the two (see Table 7).

There were two limitations placed on the selection of the random breakpoint region:

1. It could not fall within any region that is poorly characterized in the current reference genome assembly (e.g. if less than 60% of the sequence in the region was uncharacterized it would not be part of the breakpoint).
2. The breakpoint could not be on or within 1kb of any cytogenetic band boundary location. This was important as the *de novo* references used to align reads are currently generated by merging two cytogenetic bands, and therefore always have a fusion point at the band boundary. The simulated data needed to avoid a similar fusion, as this would have biased the alignment test.

The result of these limitations was that none of the breakpoints used for testing were directly within a centromere or in the Y chromosome, and very few were in telomeric regions.

Table 8 Read Simulation Parameters

READ LENGTH	INSERT SIZE MEAN	INSERT SIZE SD	COVERAGE
100	365	62	3x

Simulated parameters obtained from the TCGA sample TCGA-A3-3308-11A-01D-2048-08

Using the ART (Huang et al. 2012) Illumina read simulator reads were generated both for the normal chromosomes (e.g. FASTA files of the chromosome pairs) and then for the artificial breakpoints.

A sample BAM file was initially evaluated to obtain realistic parameters for the simulated data (e.g. read length, insert size mean, insert size SD), in this case the germline sample from a TCGA patient provided the parameters (see Table 8). A low coverage parameter (3x) was then used to generate sparse reads similar to what would be obtained from the discordant and unmapped reads post-alignment from a real sample. Once the artificial reads had been generated for the breakpoint, they were merged to the normal reads at a consistent ratio to generate a FASTQ file containing a breakpoint with known properties (the read ratio was 1:100 breakpoint to normal read). To test the MultiSieve scoring system, 20 FASTQ files containing different breakpoints were

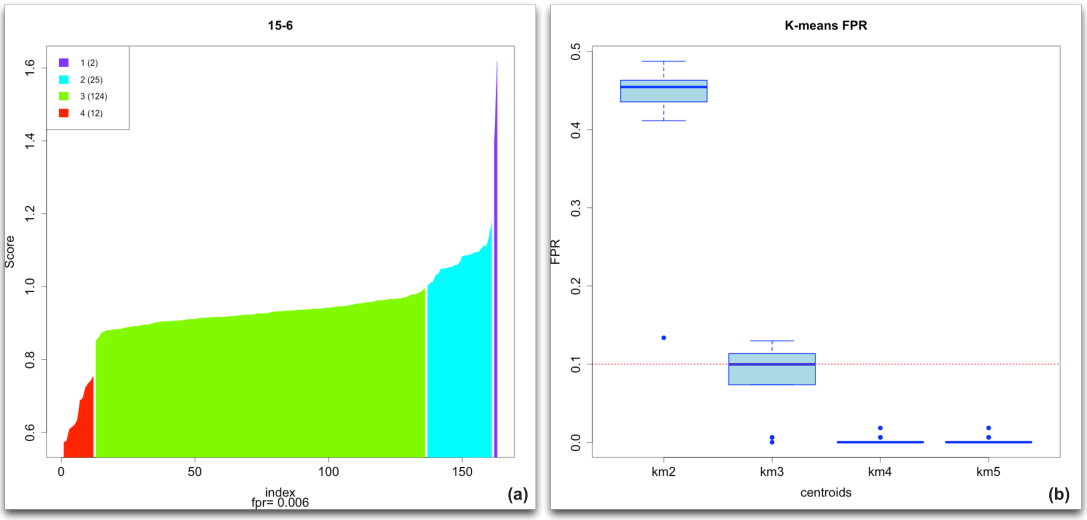
generated.

4.2.1 Sensitivity Estimation

Each of the 20 FASTQ files were in turn aligned against a bank of *de novo* references to test whether the correct breakpoint could be identified. The bank of *de novo* references consisted of 150 randomly selected breakpoints, and 20 that modeled the breakpoints introduced into each of the corresponding test FASTQ files.

In each simulated set the alignment against the *de novo* references was evaluated to obtain a $\boldsymbol{Tx}$ score (see Methods section 3.2.3). Identifying the high scoring regions for each simulation set required selecting an appropriate cutoff. This cutoff needed to be reasonably strict due to the high likelihood of erroneous alignments from lower quality reads. Specifically it was expected that when aligning the same read-pair to multiple different references, multiple reported alignments would result. As each reference alignment was independent of all others this meant that there was not one single "best mapping" alignment selected for a read-pair.

Figure 26 K-means Cluster of Tx & False Positive Rate Selection



K-means clustering of the Tx scores from the simulated datasets. In (a) one of the simulation pairs, 15-6, is shown as an example. The scores are colored according to the cluster each is a member of. The purple cluster on the far right is the cluster that includes the correct region. In (b) the false positive rates were calculated for each of the 20 simulated datasets at multiple centroids. Starting with a centroid of 4 the FPR falls below 10% (the dotted red line).

In each test the distribution of $\boldsymbol{Tx}$ scores displayed a long tail for the high values. In order to identify a cutoff that would include the correct region and include the fewest false positives k-means clustering was applied using different numbers of clusters (see Figure

26). In this test two clusters, as expected, resulted in an expected high FPR (nearly 50%). With 3 clusters some of the simulation results displayed a FPR around 10%, and at 4 clusters gave a consistent FPR of 5-7%.

Based on these simulations k-means clustering with a minimum of 4 clusters provided the best sensitivity and was then used in all subsequent analyses. The risk was in increasing the false negative rate, particularly in data that may include a high rate of structural variation or significantly heterogeneous tumors (e.g. leukemias frequently exhibit both high rates of chromothripsis and significant heterogeneity in cellular populations).

4.2.2 Coverage Parameters

While the sensitivity estimation provided a method through which to select the most likely regions that represented breakpoints in the data, one issue it did not address was the effect of read coverage for a breakpoint region. In most tumor samples a complex mix of sub-clonal populations will be present with potentially unique structural variants that will result in a low frequency of sequence reads. This required that the impact of the low frequency of reads in a sample be assessed. To understand this effect the simulated data was used.

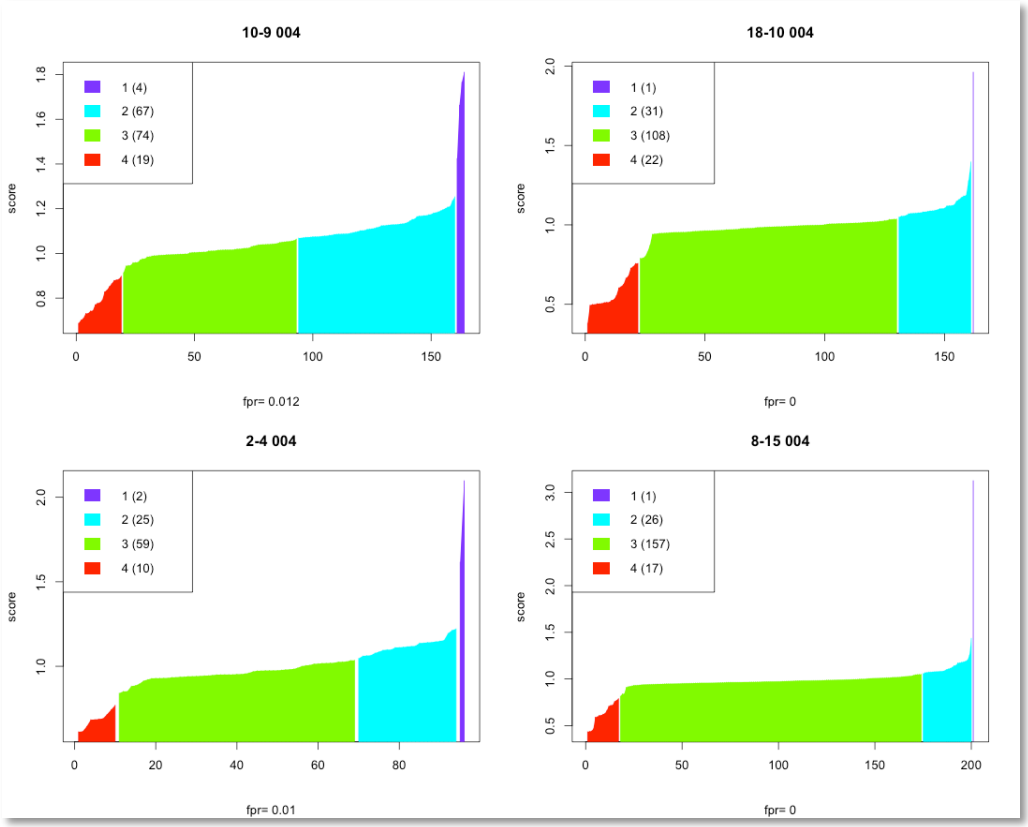
It was found that when there was high coverage (e.g. when breakpoint reads made up 1% of the reads in each simulation) the correct region was scored highly and identified in all 20 simulations. However, when the coverage around the breakpoint was lowered this resulted in lower identification rates and higher rates of false positives.

This coverage issue was overcome by applying a suitable weight to the positional clustering ratio parameter in the $\boldsymbol{Tx}$ score. This parameter was based upon clustering the aligned reads that span the breakpoint. This was tested by decreasing the coverage of breakpoint reads to 0.4% (1:250 reads) and running the same analysis in 5 simulations. In these tests only 2 of the 5 simulations were able to identify the correct structural variation, and while the FPR for these 2 regions was below 8% the other 3 could not identify the correct variant at all. Weighting the positional clustering ratio parameter proportionally to the coverage difference between the 1% and 0.4% results in correct identifications for 4 of the 5 simulations (see Figure 27) and a FPR below 5%. The weighted parameter had an effect on low coverage samples as it increased the rate of identification without altering the FPR.

In tumor samples variations will be present at various rates with inconsistent coverage due to heterogeneity of the tissue samples. With actual patient data weighting the clusters may improve identification of higher coverage breakpoints without a similar improvement for low coverage breakpoints simply due to the fact that coverage is inconsistent even

across small regions of the genome. Therefore weighting this parameter should be used cautiously and conservatively where there is no prior information available on the purity of the tumor sample or how heterogeneous the cellular populations are.

Figure 27 Weighting for Coverage



Identification can still be made with decreased coverage if the weighting parameter is increased accordingly. These plots show the k-means clusters from 4 simulations where the breakpoint reads are present at a rate of 1:250 normal reads. Weighting the clustering parameter accordingly preserves the identification and FPR.

4.3 Selecting Appropriate References

The results of the karyotype analysis offer an idea of chromosomal instability in the development of tumor cell populations. It is clear that some regions of the genome are more vulnerable to breaks (fragile) than others (e.g. due to open DNA, micro homologies, chromosome locations). This means that prior information on known fragility based on cancer-specific instability could be used directly in an analysis. Such a direct approach would allow testing specific regions for breakpoints. Using such an approach would result in selecting regions such as 21q22 and 10q23 in a prostate cancer sample (Berger et al. 2011), or anywhere in chromosome 17 in a breast cancer sample (Przybytkowski et al. 2014). However, such an approach has the distinct disadvantage of missing previously unknown, or novel variants in a patient.

An alternative to an "informed" section of putative breakpoints would be to test all possible pairwise combinations of genomic regions. Computationally this presents major problems though even if the cytogenetic regions selected are only at the level of the 'major' bands (e.g. locations of Giemsa-stained bands within the genome sequence assembly). As noted in the Methods chapter, there are 320 major bands and a pairwise combination of these results in $C_{320,2}$ (51,040) regions to be evaluated. Exhaustively evaluating all of these is computationally prohibitive (e.g. 96 years for a single sample).

Therefore a combined approach where the search is optimized based upon prior knowledge and HPC methods are adopted provides the most suitable approach to solving this problem.

4.3.1 Search Optimization

Generating all possible combinations of putative chromosomal fragile sites is infeasible due to the computational resources necessary even for a single sample. One solution to this problem is to narrow the search space for regions that have been shown previously to harbor structural variations. Due to the size of the search space a knowledge-driven optimization strategy was tested in comparison to randomly selected regions (as discussed in the Methods chapter).

Table 9 DE vs Random in Cell Line

	DE	Random
HCC1954	4q13-15q21	4q13-1p12
	17q23-4q13	16q24-4q13
	4q13-2q31	4p14-4q13
	4q13-2q14	5q34-4q13
	4q13-1q21	17q21-4q13
	3p23-9q13	15q15-4q13
	10p14-9q13	19q12-9q13
	21q21-9q13	8p21-9q13
	4q21-9q13	9q13-9q33
	4q32-9q13	
	9q13-12p13	
	Xq21-9q13	
	3q12-8p11	

Comparing the optimization algorithm to randomly selected regions finds more results in the top cluster.

As shown in the validation tests using simulated structural variants, the scoring metric found regions that were likely to contain structural variation with a 10% or less FPR when using k-means clustering to select the set of high scoring regions. To explore the validity of using prior knowledge as opposed to randomly selected regions further tests were

performed. The purpose was to demonstrate that regions selected using prior knowledge would provide more accurate results.

This was shown using samples available from TCGA. The first test involved using a whole-genome sequence for the breast cancer cell line HCC1954.G31860. The Broad Institute originally aligned it against reference GRCh37 using BWA. According to SAMtools (Li et al. 2009) 92% of the 3.1 billion reads were correctly aligned. After applying a QA procedure, which removed all reads that failed quality control or were marked as duplicates from the remaining reads, there were 119 million reads left. These unassigned reads were those where one or both reads were unmapped, or ones where the pairs had mapped to different chromosomes (discordant). These reads were then aligned against both the randomly generated regions, and those selected through the optimization algorithm with BWA.

The *de novo* references were used to align the filtered reads from the TCGA sequence data, and scored as described previously. Following this, each set (random and DE optimized) was evaluated for abundance of high scoring references. In this test the optimization algorithm (DE) identified 13 regions as being structurally variant, while the random set identified only 9 (see Table 9). Of particular interest is that in the DE regions the bands 4q13 and 9q13 are identified as part of multiple regions, suggesting that these could be part of complex rearrangements or be particularly fragile in this sample. The random set was also enriched for both of these regions. Furthermore, 4q13 was found to be enriched in a subsequent analysis of a breast cancer patient (discussed in section 4.4.1 Patient Variation Analysis).

Table 10 DE vs Random in LUAD

	DE	RANDOM
LUAD	11p15-12q11	8p21-9q13
	6q22-11q21	9q21-Xq12
	Xq21-9q13	2p14-5p12
	6q15-10p13	5p12-3q22
	21q22-14q22	12p13-9q13
	10p14-9q13	
	9p22-14q21	
	5q33-9p23	

Search optimization selection (DE) vs random selection in patient samples shows that the optimization results in a higher rate of variant identification.

A second test was also undertaken using a TCGA patient with both germline and tumor samples. These sequences were aligned against a set of *de novo* references

selected by the search optimization algorithm (278) and a randomly selected set (297). In both the random and DE sets the reported regions were those found only in the top hits from the tumor sample after filtering out hits that were also in the germline sample. In this patient the DE optimization method also resulted in identifying more regions (8) as compared to the random selection (6). Unlike the cell line above, there was no enrichment for a single region (see Table 10).

These tests showed that the DE method was able to identify more SV regions than random sampling, as it was able to provide a more diverse/representative set of breakpoints from the knowledgebase of prior information. For this reason the DE method was used as the standard method of select fragile sites for subsequent analyses.

4.4 Patient Data: Germline/Tumor pairs

To demonstrate the applicability of MultiSieve, nine matched tumor/germline patient genomes from TCGA across seven different cancer types (see Table 11) were analyzed to identify the structurally variant regions. As there was a computational resource limitation an analysis of a single cancer type across all patients (typically 100+ genomes) was not feasible. The HPC section in the Methods chapter discusses strategies proposed for allowing this type of population analysis to be performed, and the corresponding HPC results section (below) demonstrates the feasibility of using such scale-out systems. Instead, multiple different cancer types were used to evaluate how robust the method was in regards to commonly identified regions. Such regions (shared across different samples) were surmised to be due to issues with alignment artifacts or highly repetitive regions, rather than being due to common fragile sites across all cancers. With large-scale structural variation there is little reason to expect many commonly shared variants, as most of these variants are acquired late in a clonal population's evolution. Therefore, in order to limit the introduction of additional bias a new set of *de novo* references was not generated for each patient, rather each sample was run against the same set of sequences representing the same likely fragile sites in the human genome. For this analysis the optimization algorithm was used to create 276 unique *de novo* references (the optimizer starts with a population of 200-400 and generated no fewer than 200 references). Each of the patient samples was filtered using SAMtools for all discordant, unmapped, or partially unmapped reads. The reads were aligned against each *de novo* reference as described previously, and the alignment was evaluated to obtain the $\boldsymbol{Tx}$ score.

In each patient the top cluster as determined with k-means using 4 clusters, was obtained. In the final step the regions in the top cluster from the germline (e.g. the presumed "normal") sample were filtered out of the regions from the tumor sample. This

was performed specifically because large-scale structural variations in tumors are somatic mutations. This was especially important as unbalanced translocations are common in tumors, but developmentally lethal so will not be found in the germline. Each patient was also analyzed with the commonly used reference-based (see Introduction) tool, BreakDancer. Finally, one patient's variants were evaluated for known oncogenes or pathways. These results were reported in Killcoyne and del Sol, 2015.

Table 11 TCGA Patient Samples

CODE	DISEASE	TCGA BARCODE	SAMPLE TYPE
BRCA (1)	Breast invasive carcinoma	TCGA-BH-A0DK-01A-21D-A060-02	Primary Solid Tumor
		TCGA-BH-A0DK-10A-01D-A060-02	Blood Derived Normal
BRCA (2)	Breast invasive carcinoma	TCGA-A1-A0SM-01A-11D-A19H-09	Primary Solid Tumor
		TCGA-A1-A0SM-10A-02D-A099-09	Blood Derived Normal
COAD (1)	Colon & rectal adenocarcinoma	TCGA-QG-A5Z1-01A-11D-A28G-10	Primary Solid Tumor
		TCGA-QG-A5Z1-10A-01D-A28G-10	Blood Derived Normal
COAD (2)	Colon & rectal adenocarcinoma	TCGA-AZ-4315-01A-01W-1461-10	Primary Solid Tumor
		TCGA-AZ-4315-10A-01W-1461-10	Blood Derived Normal
GBM	Glioblastoma multiforme	TCGA-02-2483-01A-01D-1494-08	Primary Solid Tumor
		TCGA-02-2483-10A-01D-1494-08	Blood Derived Normal
KIRC	Kidney renal clear-cell carcinoma	TCGA-A3-3308-01A-01D-2094-10	Primary Solid Tumor
		TCGA-A3-3308-11A-01D-2048-08	Solid Tissue Normal
LAML	Acute myeloid leukemia	TCGA-AB-2905-03A-01D-0739-09	Blood Derived Cancer
		TCGA-AB-2905-11A-01D-0739-09	Blood Derived Normal
LUAD	Lung adenocarcinoma	TCGA-05-4384-01A-01D-1751-02	Primary Solid Tumor
		TCGA-05-4384-10A-01D-1751-02	Blood Derived Normal
OV	Ovarian serous cystadenocarcinoma	TCGA-04-1331-01A-01D-A324-10	Primary Solid Tumor
		TCGA-04-1331-10A-01D-A324-10	Blood Derived Normal

This lists the patients downloaded from TCGA for analysis. The shaded cells are the germline samples belonging to each patient pair. Note that the TCGA combined samples for colon and rectal cancers in the COAD set.

4.4.1 Patient Variation Analysis Results

The results can be broken down into three classes: patients with no structural variations found; patients with low numbers of variations, defined as 5 or fewer; and patients that are highly variant with more than 5 regions found.

No Structural Variants Identified

In two patients the final list of structural variants was empty. The first patient was a case of glioblastoma multiforme (GBM). Both the germline and tumor analyses resulted in very few regions being selected in the top clusters, and all were shared between them. However, this is not surprising as GBM is not a tumor that is known for high rates of genomic instability the way the leukemias are. Recent RNA-seq analysis on the GBM cohort in TCGA found that while 53% of the samples exhibited gene fusions the major hot spots for these were in 7p11 and 12q14-15 (Shah et al. 2013). However, they could not identify genomic breakpoints in these regions and concluded that these were primarily copy number variants. The cohort analysis of the GBM patient included in the MultiSieve analysis (see Table 11) found only one intra-chromosomal fusion variant within 10q26. MultiSieve tested one pair of *de novo* references that included 10q26, but it was scored in the third cluster due to a very low positional cluster weight, meaning very few reads were found which could indicate a breakpoint in this region. In addition, none of the region pairs that included 12q14-15 or 7p11 were identified in either the normal or tumor sample in the top cluster.

The second sample with no variations identified was one of the patients (TCGA-AZ-4315-01A-01W) with colon rectal adenocarcinoma (COAD (2)). Note that this class of tumors included two types (or sub-types) and that these were treated as a single cancer type by TCGA. That there were no structural variants in this patient was consistent with both the stage (IIA – no metastasis, no lymph node involvement) and with the previous analysis reported by TCGA (The Cancer Genome Atlas Network 2012). This sample was not associated with any small-scale structural variation, copy number variation, significant numbers of small variants, or even any pathway alterations.

Low Structural Variation

Four patients exhibited low structural variation, with 5 regions or fewer identified as highly scoring and not found in the high scores of the germline sample. Two of the patients had a single region, different in each case, identified as being highly scored for structural variance (see Table 12).

Table 12 Low SV Patients

BRCA (1)	BRCA (2)	KIRC	COAD (1)
Xq21-9q13	17q23-4q13	3q12-8p11	3q12-8p11
	4q13-2q14		11p15-12q11
			10p14-9q13
			7p11-13q14
			Xq21-9q13

Patients with low rates of structural variation as identified by *de novo* references.

The second BRCA patient had two regions identified as being structurally variant, both of which included 4q13 suggesting that this band may be driving the scores for both. This region was also found to be overrepresented in the breast cancer cell line analysis (HCC1954.G31860) and contains a number of genes implicated in breast cancer including EREG (epidermal growth factor receptor ligand epiregulin), which is associated with breast cancer metastasis specifically (Eltarhouny et al. 2008). This belongs to the class of Erb/HER ligands that are expressed in many breast cancers and are related to tumor aggressiveness. An additional 10 regions that included 4q13 were generated and tested along with the original 276 regions. All are found in the highest scoring cluster in the tumor, but not in the normal sample (see Table 13). This pattern could indicate a copy number mutation in the region instead of a translocation.

Table 13 Additional BRCA Regions

BRCA (2)	
4q13-1p12	15q15-4q13
16q24-4q13	17q21-4q13
17q23-4q13	5q34-4q13
4p14-4q13	**4q13-2q14**
4q13-15q21	4q13-2q31
4q13-9p24	4q13-1q21

All 12 regions that include 4q13 are highly scored in the tumor but not the normal sample for this BRCA patient. The two in red were found in the initial analysis

The fourth patient was from the COAD cohort in TCGA (TCGA-QG-A5Z1-01A-11D), and had 5 regions highly scored. This patient was sequenced after the original TCGA paper on colon and rectal cancers and so cannot be referenced directly. However, in contrast to the COAD patient with no variants, this patient is at a much later disease stage (IIIB - lymph node involvement and is spread more widely through the colon walls, but has not yet metastasized). It is clear from the TCGA paper on the earlier COAD cohort that colon cancers can be split into two genomic sub-types: hypermutated with MSI; and

microsatellite stable, but chromosomally unstable, including whole arm gains and losses. One of the regions identified as variant by MultiSieve includes 11p15, which was found to be one of the most common amplifications in the chromosomally unstable subtype of tumors in the cohort analyzed originally by TCGA.

High Structural Variation

The final set of patients exhibited high structural variation. These had more than 5 regions identified and included the lung adenocarcinoma (LUAD), ovarian serous cystadenocarcinoma (OV), and acute myeloid leukemia (LAML) samples (see Table 14).

Due to the high representation of leukemia in the karyotype data, used to generate the probabilistic pairs of regions to test, it might have been expected that either of bands involved in the Philadelphia chromosome be highly scored in LAML. None of the 6 pairs of regions, which included 9q34 (22q11 was not in any of the 276 pairs generated with the search optimization algorithm), scored highly enough to be found in the top cluster for either the normal or tumor sample. However, while it is the most common translocation in chronic myeloid leukemia (CML) it is significantly less common in AML cases.

Table 14 High SV Patients

LUAD	OV	LAML
11p15-12q11	3q12-8p11	4q23-16q13
6q22-11q21	10p14-9q13	3q12-8p11
Xq21-9q13	17q23-4q13	2q13-14p13
6q15-10p13	22q13-9q13	3q27-6q15
21q22-14q22	5q32-9q13	17p11-13p12
10p14-9q13	4q13-2q14	5q32-9q13
9p22-14q21		
5q33-9p23		

Patients with high rates of structural variation include lung, ovarian, and leukemia.

The LUAD patient had the largest number of highly scoring regions found after removing those found in the germline sample, with 9. The most salient feature of these regions in comparison to those found in the OV patient, is that there is very little overlap between the regions in regards to shared bands. This pattern suggests a high level of instability across the genome in this patient, rather than a single region that breaks and recombines multiple times. Such a pattern would be consistent with the hypermutation characteristics in tobacco related lung cancers, and this patient was reported to have smoked tobacco for 20 years.

4.4.2 Comparison to Reference-Based Method: BreakDancer

All of the patients were also analyzed with BreakDancer, as this tool continues to be the most commonly used and most commonly compared to, structural variation tool particularly with regards to large-scale variations.

BreakDancer reports translocations based on clusters of discordantly aligned reads in a given chromosomal position. It uses only those reads that aligned and were reported in the original BAM file and so uses significantly less information than MultiSieve. Due to this large difference in the quantity and specific type of data being used, it was unlikely that there would be much overlap between what was found using the MultiSieve method and BreakDancer, and none was found. However, some very important differences between the two methods are highlighted by this comparison (see Table 15).

The first difference is that BreakDancer identified and highly scored translocations in the same chromosomal regions across all the patients and tumor types. Across all 32 regions that were reported, 26 were found in at least two patients. For instance, translocations in the 1p11-17p11 region pair were identified in 6 of the 9 patients, and translocations in 1p11-11p11 were found in 4 of the 9. In the COAD patient where the multiple *de novo* references method found no difference between tumor and germline, BreakDancer identified only a single pair (3q27-6q15) that it did not also find in the other patients. This region pair was tested by the MultiSieve method as well, but was not highly scored.

Related to the common identification of translocation regions, BreakDancer also identified centromeres more frequently than the MultiSieve method. Across all of the samples that were analyzed, centromeres were significantly overrepresented in BreakDancer's results with 27 of the 32 chromosomal translocations including at least one centromere in the regions represented by the aligned reads. However, centromeres are poorly characterized across most of the chromosomes due to their highly repetitive sequences (Treangen and Salzberg 2012). The centromere (1p11) that was found in most of the translocations across samples is an example of this, with 80% of the bases lacking a known assembly. This results in sequence reads which have multiple correct alignments throughout the genome, but that have been reported as aligning only once due to the limitations on alignment algorithms.

Finally, as BreakDancer is limited by the read alignments reported in the original BAM file it also reports erroneous translocations due to incorrect alignments. The clearest example of this is that two translocations in the top scored variants for one of the BRCA patients include Yq11. The clinical data associated with this patient lists them as female, and while it is not impossible for the patient to carry a Y chromosome, it is unlikely as that

generally results in other diseases that would have been noted in the clinical data or prevented the patient from being included in the study. The issue is most likely one of alignment. Generally the Y chromosome is not removed from the reference index prior to aligning sequence reads, providing a good example of the need for more patient-specific alignment.

The fact that the chromosomal regions found tended to be shared across patients and cancer types strongly suggests that only very common structural variations are found by BreakDancer. However, this is primarily due to the fact that BreakDancer and other similar tools rely on the original alignment. As alignment algorithms generally report only a single alignment (due to computational limitations) tools such as BreakDancer that rely on the primary alignment information are left with extremely limited information to make identifications with. This issue is compounded by the complexity of tumor samples where multiple clonal or sub-clonal cellular populations may be present with a mixture of high and low mutation genomes.

MultiSieve, as it realigns against selected *de novo* references, does not have these limitations.

Table 15 BreakDancer Results

BRCA (1)	BRCA (2)	COAD (1)	COAD (2)	GBM
1p11-19q11	1q12-Yq11	1q12-Yq11	1q12-21p11	**1p11-17p11**
	1q21-Yq11	1q43-10p11	1p34-6p22	
	1p11-17p11	1q21-Yq11	1q21-16p11	
	1q21-4p11	1q21-4p11	**1p11-17p11**	
	1q21-16p11	1q21-21p11	3q27-6q15	
			1p11-11p11	
KIRC	**LUAD**	**OV**	**LAML**	
1p11-17p11	**1p11-19q11**	1p22-17q12	**1p11-17p11**	
1q21-4p11	**1p11-6p11**	1p22-9q22	**1p11-11p11**	
1q12-21p11		**1p11-11p11**		
1p11-11p11		1q21-16p11		
		1p11-17p11		
		1p34-6p22		

This table shows the results for the same patients run in BreakDancer, a commonly used structural variation identification tool and typically the only one used for large-scale variants. Bolded regions include the centromere 1p11, which is a poorly assembled region. The two regions highlighted in red are incorrect calls that are due to alignment errors as they include the Y chromosome in a female patient.

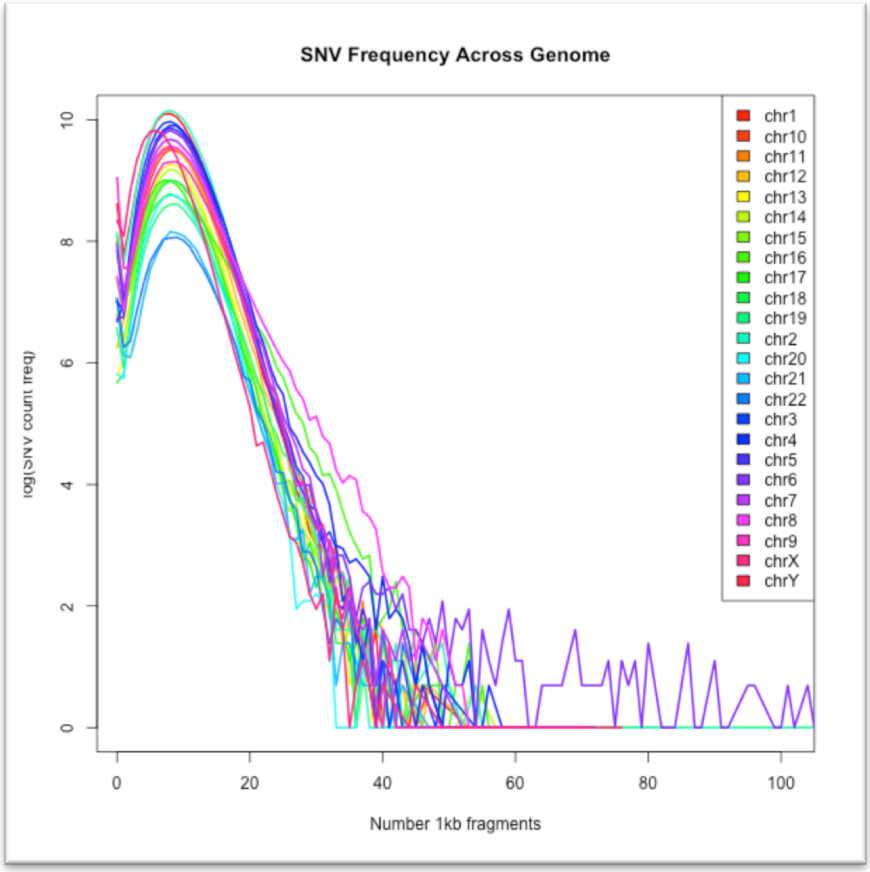
4.5 Population Scale Analysis

A limitation of the MultiSieve, and all other SV methods, is the significant computational resources required. Even with the optimization procedures implemented, MultiSieve analyses were only able to be undertaken on small numbers of genomes. The methods developed for scale-out, to alleviate this problem, are discussed in the Methods chapter, and the results the analyses undertaken using these frameworks is discussed for completeness below.

4.5.1 Variation Frequency Analysis

An analysis to determine the variation frequency required that all of the variants be from a source that was not specifically examining genomic diseases. Therefore, the Ensembl service was used to access variant information for all small variations that were tagged as “validated” in either 1000Genomes or HapMap. This information was used to determine the frequency of mutation as these are from ‘normal’ (not cancer) populations.

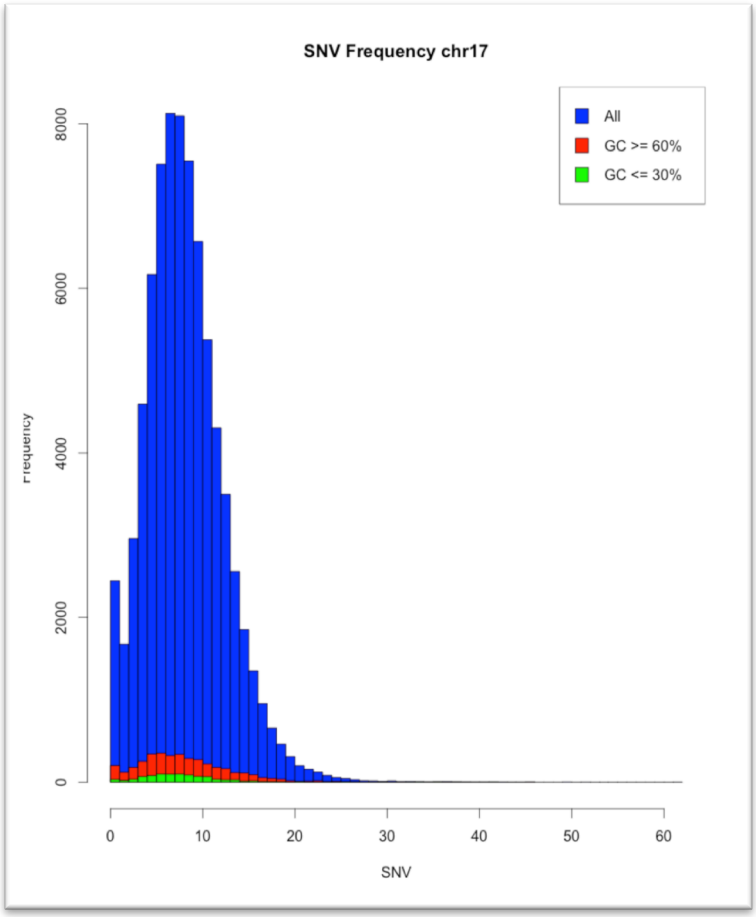
Figure 28 SNV Frequency per Chromosome



The mean frequency is nearly identical across chromosomes The “jitters” at the extreme end of the distributions are an artifact due to the rarity of fragments with higher numbers of SNVs.

GVF files were parsed to select variants that included the 'validation' or 'evidence' tag in the features. Six classes of variants were identified. All of the class definitions are alterations as compared to the reference genome: deletions, one or more contiguous nucleotides have been removed; insertion, one or more contiguous nucleotides as been added; SNV, single nucleotides where different nucleotides are present; substitutions, the base pair length of the alteration is the same as the reference but the sequence has changed; tandem repeat, adjacent regions have been copied more than once; sequence alterations, an uncharacterized alteration to the genomic sequence (e.g. SNV, indel, insertion, inversion, etc).

Figure 29 GC/SNV correlations



There was no correlation between the frequencies of SNV directly with GC content. However, when comparing to high/low GC content weak correlations appear. Red bars show high GC (cor = -0.20). Green bars show low GC (cor = -0.18).

The locations, variant class, and reference/variation sequences were preserved on a per-chromosome basis. Each chromosome was then broken into 1kb segments, the variations that were defined in each segment were used to create several profiles that included: the frequency of each variant class in the segment; the specific base pair polymorphisms; and structural information (e.g. coding region coverage, GC content,

methylation predictions).

A total of $2.5x10^7$ variants were validated in 1000Genomes and HapMap and 99% of these were SNVs (25,636,921). Sequence alterations (5,959), deletions (2,614), and insertions (2,614) were between 0.01% and 0.02% of the variations, while substitutions were rare with only 29 instances identified, and only a single tandem repeat was found. Due to the comparative rarity of all other variants most of the analyses for the variant frequency were focused on SNVs.

The distribution of the SNV frequency was highly similar across all chromosomes with the mean around 14 for each (see Figure 28). The frequency of fragments having a high number of SNVs (e.g. more than 40) drops quickly, though there are a few fragments with as many as 200 SNVs.

However, this does show some difference, particularly as the size of the chromosomes decreases so each subsequent analysis was performed per-chromosome rather than on a genome-wide scale.

In order to determine under what conditions these variants might occur three structural properties of the genome were investigated:

- *Coding regions* may be related to genome or chromosome-wide variation as variation occurring near coding sites could be less likely to occur due to protein function impairment or a more robust DNA repair mechanism in these regions to preserve function (Lercher and Hurst 2002).
- *Predicted methylation of CpG islands* (CpGI) was considered as methylated regions have been shown to be common mutation sites, therefore frequencies could be increased in regions where CpGI are found.
- *GC content*, while related to both CpGI and coding regions is a more general profile, and evolutionarily GC-rich regions have been shown to undergo higher rates of gene conversion (Kudla, Helwak, and Lipinski 2004; Katzman et al. 2011).

Coding regions were tested first by mapping the 1kb fragments to known coding regions (defined by locations in Ensembl), however no correlation between the frequency of the SNVs and the coding/non fragments was identified. The same was true when exons and introns were mapped to the fragments. Methylation was investigated by using predicted methylation sites generated by HDMFinder (Das et al. 2006). This was used to compare the fragments that fell within the highest frequency (e.g. 2-28 SNV per 1kb) with the predicted methylation regions. A hypergeometric test for methylation applied for both CpGI and non-CpGI methylation found no connection between methylation sites and variation frequency. This is possibly due to a bias in the specific methylation data provided to the HDMFinder tool in the initial published predictions (derived from brain tissue only). The methylated sequences were most often found near centromere boundaries and other

repetitive sequence boundaries regardless of CpGIs within those regions. In fact, the GC content of these regions was the highest predictor of both methylation and CpG content.

Independent of CpGIs the GC content of the fragments was tested against the frequency of SNVs. All of the chromosomes had a stable average GC content between 38-48% of the total nucleotides, and the 1kb fragments ranged from 8.5% to 85% GC content. Due to this wide variation there was no direct correlation between frequency and GC. Breaking the down the fragments by GC into higher and lower content did find a weak correlation however. GC content of more than 60% or less than 30% showed a weak inverse correlation with the frequency of SNVs in those fragments (see Figure 29). Based on this each chromosome was broken into bins based on GC content and each fragment assigned to a bin.

The same analysis was performed on cancer variations as validated in the COSMIC database and from all publicly available TCGA mutation files (as of January 2013). However, TCGA variants were only reported if they were not found in the matching normal samples resulting in mutation files that included only cancer mutations. So while the frequency of variations in COSMIC data was higher than in the 1000Genomes data, as would be expected, the frequency in TCGA was not. Despite this, the distribution of variant frequency was the same. The GC, exon, and CpGI analysis was performed on the COSMIC variants to find any cancer-specific patterns. However, the GC content continued to be the only structural correlation with variation frequency.

4.5.2 Small Variant Simulation Using MapReduce

Using MapReduce jobs in FIGG (Killcoyne and del Sol 2014) built to simulate genomes based on the observed variant frequencies, six whole genomes were simulated. Three of these used the frequency distribution based on the 'normal' genomes from 1000Genomes and HapMap, the other three used the frequency distribution based on variants analyzed from COSMIC and DGVa (in cancer samples), and one of those three 'cancer' samples also included a common structural variation. It is of note that the cancer variation frequencies include significantly more deletions, insertions, and substitutions than the 'normal' data and that these were much larger (several hundred base-pairs).

Both the 'normal' and 'cancer' genomes preserved the frequency distribution of their respective background data, but differed on a per-fragment count and size of variants. The simulation was not intended to replicate the exact frequency of variants found in fragment 1017 of chromosome 1 for instance. Instead it determines the GC content of the fragment, randomly selects a profile matching that GC content, and then generates the variants at random locations within the fragment itself. Nucleotide specific probabilities for SNVs and size-dependent probabilities for small structural alterations ensure that the

applied mutations are still within observed boundaries. Further details are in Killcoyne and del Sol, 2014.

To test that these genomes differed as expected they were used as reference genomes to align a 1000Genomes sample which is expected to be normal: ERX000272. Therefore the alignment to the normal reference and the three 'normal' simulated genomes should be highly similar, while the three 'cancer' genomes would be expected to have poorer alignment overall. The simulated genomes aligned as expected, with the three 'normal' genomes mapping between 95-98% of all reads with 1% or fewer orphaned reads (e.g. "singletons"), and the three 'cancer' variant genomes aligning at 90% or less (down to 88% for the genome containing a single SV) with 2.8% orphaned reads (see Table 16). This is due to the higher size and frequency of variants applied to the 'cancer' genomes. TCGA genomes that are from known cancer samples show similar mapping statistics with the rates of mapped reads ranging from 65-95% and orphaned reads ranging from 1-6%.

Table 16 Alignment statistics for each genome

	SAMTOOLS FLAGSTAT		
	Mapped	*Singletons*	*% Difference from GRCh37*
GRCh37	**98.22%**	**0.85%**	*NA*
S1	97.89%	1.00%	*0.33%*
S2	95.46%	1.09%	*2.76%*
S3	97.89%	0.99%	*0.33%*
S4H	90.09%	2.89%	*8.13%*
S5H	90.35%	2.84%	*7.87%*
S6SV	88.16%	2.88%	*10.06%*

A comparison of the 1000Genomes reads for ERX000272 mapped against each genome. GRCh37 is the current reference genome. S1, S2 and S3 are genomes generated based on normal variation data. S4H, S5H, and S6SV were generated with cancer variation data and S6SV includes a deletion of 19q. The table columns are statistics provided by SAMtools flagstat: *Mapped* provides the total percentage of reads that mapped to the genome on the left; *Singletons* provides the percentage of reads that were orphaned in the alignment.

Generating these simulations would be too slow on a single commodity computer. Breaking the genome down into 1kb fragments still leaves $3x10^6$ sequence fragments to mutate for each simulated genome. This is where the initial benefits of a highly distributed computational framework are clear. In order to generate artificial data models of genome sequences for subsequent analyses the sequence generation had to rapidly output new sequences, as the computational time required to align and analyze each is already significant. Using the Apache Hadoop (The Apache Software Foundation) MapReduce framework on the Amazon EC2 platform the data was generated rapidly and scaled nearly

linearly with the number of cores (nodes) available. Additional performance gains in the HBase implementation of storage of the new genome fragments are still possible as well as HBase set up and access was the single most significant time used in each genome test.

4.6 Chapter Summary

This chapter described the results from the analysis of breakpoint frequencies in karyotype level data, including general chromosomal instability and an investigation of the high-level regional influence on that instability. The results of the karyotype analysis were then used to generate the sequence references with breakpoints for two subsequent sections of the MultiSieve method: the optimized search algorithm that was used to select references; and the selection and generation of references for analysis of patient genomes from TCGA.

Before using MultiSieve to analyze patient data from TCGA, validation was performed using synthetically generated reads from normal chromosomes with reads around a fusion point added at a known coverage. After aligning these to a set of references that included the correct SV reference, an estimate of the false positive rate was made for cutoff selection of the $\boldsymbol{Tx}$ scores.

This cutoff was used to test both the optimization search algorithm versus random selection of references, and the patient analysis of TCGA genomes. In the patient data analysis the results were compared to the most commonly used large-scale SV detection tool to show that MultiSieve is able to identify regions of structural variation that would otherwise be missed due to the reliance on alignment to a single reference. Finally, the test of HPC approaches to enable this method to scale-out to the level of populations of genomes was discussed.

In the Discussion chapter these results are considered in context to the issues of tumor heterogeneity and chromothriptic events that continue to complicate the identification of mutations in sequence data. It also outlines several of the directions in which this work could continue to provide both a method that scales appropriately, and integrates the many other types of data that may provide information about clinically relevant structural variations.

CHAPTER 5 DISCUSSION & PERSPECTIVES

Perhaps the most important understanding gained from the massive efforts to sequence cancer over the past several years have been just how complex the disease is at the genomic level. While in a broad view the "hallmarks of cancer" are common between the different cancers, it appears that the different types share general characteristics and differences much the same any two way related species do. In other words general characteristics may be highly similar, but the existing differences are highly significant. The many large-scale sequencing efforts (e.g. TCGA, ICGC) and the smaller targeted sequencing projects have made those differences even more obvious. For instance, tumors from the same tissues in different patients do not necessarily start from the same mutational drivers or even share similar prognoses: a mutated BRCA1 is not so clearly predictive of tumor progression in ovarian cancer as was assumed; common mutations to the RAS pathway in colon cancer are predictive only if the mutation occurred in the correct sequence in an individual tumor; and some glioblastoma tumors carry specific mutations that are more commonly found in ovarian than other glioblastoma malignancies (e.g. BRCA1). In short, cancer is heterogeneous at every level from the primary tissue types, between patients, and within an individual tumor.

Recognition of the diversity of cancer led to significant changes to cancer therapies. Where just a couple decades ago the only therapies available were extremely toxic and non-specific, today targeted therapies that interfere with molecules specific to cancer are more common. Such therapies include drugs targeting protein products that result from specific gene mutations (e.g. Gleevec), immunotherapies that appear to restore immune targeting of cancer cells (e.g. anti-PD 1 antibodies), and for cell surface receptors that are expressed in specific cancers (e.g. HER-2). Despite the focus on finding single genes and single mutations that could be directly targeted to alter any tumor cell state, all of these therapies are targeted to patients whose disease presents in a specific way. Their tumors happen to carry a mutation, or alterations to one of the 'hallmark' pathways, that make their disease sensitive to a specific inhibitor, monoclonal antibody, or primed T-cells.

The success of one of these targeted therapies, Gleevec, prompted a search for similar protein targets in other cancers. This specific target is the result of a structural aberration where regions of two different chromosomes fuse together, creating a fusion gene and protein (BCR-Abl). This translocation was identified several decades ago in karyotypes rather than in sequencing data and many other such chromosomal aberrations have since been identified in karyotypes. With the introduction of HTS the search for other large-scale aberrations and gene fusions moved from image-based methods to

sequences.

Massively parallel short-read HTS revolutionized the search for genomic causes of various diseases. With it, both population and individual variation has been more clearly observed and many new links between the genome and diseases have been elucidated. Understanding the genomic causes of cancer has been one area that has been directly and radically changed by the availability of whole genome scale data. In the last few years thousands of mutations and hundreds more genes have been identified as potential drivers of cancer progression. HTS continued the search for single genes or single mutations that could explain cancer, or offer new therapeutic targets. It also offered the potential to better understand what role the massive structural changes found in karyotypes could play in the development and progression of the disease.

This search has been limited by the structure of the data as well as the heterogeneous nature of tumor tissue samples. Current HTS technologies rely on sequencing the ends of a DNA fragment that is, at most, 1000 bp in length (often much shorter, only 400-600bp), and typically only 100-200 bp of each end of that fragment are sequenced. In large genomes (e.g. mouse, human) a single sequencing run will produce millions or billions of reads for a single genome. These reads are then aligned to a reference genome in order to locate the original fragment in the genome, however, finding a unique alignment against a reference that is billions of base-pairs in length is a computationally intensive and complex task.

Evidence for a breakpoint in the genome that results in a SV relies on the few reads that span either side of the breakpoint. This is complicated as the regions surrounding breakpoints are highly likely to contain microhomologies that result from DSBs, or tandems repeats (e.g. mini/microsatellites), and are often indicators of fragile regions. These repetitive regions are difficult to sequence and are more prone to errors, especially if they are GC-rich. The reads generated in these regions are more likely to be highly similar to multiple locations in the genome as well. In tumors the thousands of small variants and the high potential that large-scale rearrangements and chromosomal duplications have occurred magnify the problem. The result is that tumor sequences often have poor alignments, with some highly mutated genomes aligning only 60% of the generated reads (alignments from a ‘normal’ genome are around 98%). Current methods for identifying SVs in sequence data rely on reads that have aligned, even if they aligned poorly. Due to the high concentration of repeats around a breakpoint the reported alignment is not the only, or even the most likely, correct aligned position.

The poor alignments resulting from these regions have generally been the only evidence available to breakpoint detection methods. It has only been in the last three years that researchers have begun to look beyond the standard reference template for

answers. This has led to reexamining the use of computationally intensive methods for *de novo* alignment that are appropriate for targeted identification, and the development of methods that focus on comparing reads between two samples directly. These new methods use the standard reference at various points in the analysis, but typically not in the initial analysis. In other words, the use of the reference genome template itself is the constraint in analyzing SVs.

Heterogeneity in the tumor samples from multiple sub-clonal tumor populations, adjacent tissue, and normal cells further complicate this process. SVs (and other mutations) that could be drivers for clonal expansion following drug treatment or metastasis are likely to be found in small cellular populations within the tumor. When mixed in a single genomic sample the reads from a sub-clonal SV are likely to be low coverage due both to the small population and to the high potential that surrounding sequences are repetitive. Single-cell analysis has been proposed to overcome this issue, but questions regarding how many cells would be needed and how many regions of the tumor should be sampled have not been answered. This would also effectively increase the data scale for each patient geometrically as a single sample from the tumor would no longer be enough.

With current technology, both computational and short-read sequencing, identifying SVs using available data (tens of thousands of sequenced patients) is necessary. Altering the reference genome directly and including the previously unmapped reads enables identification of multiple regions for SV detection.

5.1 In Silico References

This project reformulated the problem of SV detection in short-read sequencing by not assuming that the reference was correct, but instead by using many references that model SV regions directly. Modeling potential SVs directly in the reference enabled the use of existing tools in an innovative method to select the models that best represent real SVs. While this still used a reference, that reference template is no longer treated as the final answer prior to investigating the question of whether or which SVs may be in the tumor.

5.1.1 Advantages of this approach

This approach offers a number of important advantages in the identification of large-scale structural variants. First, by using available biological knowledge about chromosomal aberrations an appropriate search space can be defined. Karyotype information has long been used to provide clinicians and researchers with important prognostic information about a tumor. Using the same information clinicians use defines a new search space for SV detection while still allowing that space to include aberrations

that may not be frequently identified in the public data sets. Furthermore, the search space can be targeted if karyotype information is already available about a specific patient. References can be generated to conform to what is known about the karyotype and read alignment can be investigated through a tree of related references. For instance, if a FISH/SKY karyotype has been visualized and a derivative chromosome that includes sections of chromosomes 4,17, and X is found, a set of references focused on regions in those three chromosomes can be directly investigated and the most likely alignments selected.

Secondly, by altering the references provided to the alignment algorithms the inherent computational limitations of the alignment can be avoided. A read can be reported for multiple possible locations, which is especially important in regions that are likely to be repetitive and have a low frequency of supporting reads as is typical of breakpoints. Utilization of HPC technologies that have become both easier to use, and widespread in sequence analysis allow the selection and detection to efficiently scale to the available resources. Scaling the search effectively provides information on genomic locations that the standard methods cannot.

Finally, while the initial knowledgebase from karyotypes is not cancer-specific due to the availability of data, as more information regarding breakpoints becomes available this resource can be improved. Specifically with more information on breakpoints found either within a single cancer-type, such as gene fusions that are commonly found in colon cancer or glioblastoma, the knowledgebase can be targeted. Both cancer-specific and pan-cancer knowledgebases could be used to generate SV reference models that are closer to the patient sequence while still providing the space to search for less common SVs. This would also further speed up the analysis process, as the targeted knowledgebase could cache some of the references directly rather than generating them each time, cutting down on one part of the computational time required.

5.1.2 Limitations

Increasing the number of references, even using a knowledgebase to generate more targeted references, also increases the search space and the incidence of non-unique alignments. Increasing the search space was a necessary condition to enable identification in mixed samples, and one of the primary aims of this method. However, it also directly increases the number of regions reported for a single read since multiple alignments can be reported for each read. Each read has a greater potential to align to multiple locations simply due to the independent alignment of each reference, the alignment algorithms are no longer being tasked with selecting the ‘best’ mapping. While the result of this is that each read could have multiple alignments, it is not necessarily

incorrect. The low frequency of reads around a breakpoint and the high likelihood that the surrounding sequence may be repetitive means that these reads will have multiple mappings with no clear answer as to which is correct. The standard approaches will select one location alignment at random when multiple alignments are found, while this approach simply reports all alignments allowing the single read to provide only one piece of evidence for a breakpoint.

Computationally this approach is resource intensive, requiring at least a general cluster environment and preferably a framework that allows for the use of MapReduce. Even using HPC frameworks and cluster environments there are more possible regions than can be reasonably tested in each genome. This is partially overcome through the use of the prior knowledge of large-scale aberrations and a search optimization algorithm, as well as limiting the reads used to only those that are discordant and partially/fully unmapped. However, these optimizations limit the search space that can be tested in each genome and as it cannot search all possible regions some SVs will be missed.

5.2 Future work/Outlook

While we have a few examples of large-scale structural variation resulting in a fusion event between two genes that form a mutant protein with direct consequences for a patient, such examples are relatively rare. As the extent of chromothripsis in tumor genomes has made clear that significant alterations to genomic structure is survivable at the cellular level, it has also become clear that their effect is not always directly observable. In some cancers SVs are linked with poorer outcomes (e.g. IGH translocations in multiple myeloma) without an obvious biomarker, while some result in a therapeutic target (e.g. BRC-Abl in CML, PML-RARA in acute promyelocytic leukemia, EWSR1-FLI1 in Ewing sarcoma). These targets have enabled the development of some of the most effective therapeutics, sometimes dramatically reducing mortality. In recent years structural rearrangements that lead to gene fusions have been found across multiple types of cancers and tissues, but few of these have also been clearly associated with a clinical outcome or biomarker.

In fact one of difficulties in structural variation analysis in cancer has been that not all SVs result in a directly gene related product or gene fusion event. The breakage and recombination of an intra- or inter-chromosomal translocated segment could occur in intergenic regions, affecting gene regulation or transcriptional enhancement rather than directly inactivating the gene or increasing overall expression. The heterogeneous cell populations within a tumor may also be regulating each other, making a SV or a fusion that occurs in a sub-clonal population undetectable in current expression studies. Identifying these fusions could provide clues as to the future progression of the tumor, or

eventual metastatic populations. In order to do this, future investigations will need to integrate multiple levels of data.

Additionally, as the coverage in short-read data increases the need for scalable analyses becomes more important. Large compute clusters have become standard equipment for a sequencing project and any new methods must be capable of using them effectively.

5.2.1 Scaling Up

This method was designed to take advantage of HPC technologies in the generation and alignment of hundreds of small references. This initial design was intended only for single genomes by generating hundreds or thousands of small references that model SV regions and analyzing them simultaneously. However, with thousands of cancer genomes already available and smaller projects generating tens or hundreds more for a single analysis it is also becoming necessary to analyze genomes at the population level. In order to do this efficiently greater use of HPC technologies will be important.

Reference Caching

In the analysis prepared for this thesis each individual genome was compared against the same set of nearly 300 small references. This was specifically done to compare the individual differences in identified SVs, however if the analysis had been for a population of genomes that were all from the same cancer type (e.g. ovarian patients) or single cell analysis from the same tumor (e.g. 10 different cell samples) it would be sensible to also use the same set of references for alignment.

A file cache with a rapidly searchable index of all possible major SV references and their attendant alignment index would enable all of the generation time to be upfront. In this case the cache would be generated at one time, when a new genome build is released, or simply added to with each new set of DE generated references. This would require that a minimum of 2.5 TB be set aside for this cache, even if the cache were generated dynamically with each new DE selection. Additionally, as the combinatorial number of major bands is significant, (e.g. $C_{320,2}$ or 51,040) the file index and organization of the reference files themselves would need to be generated such that the references are rapidly accessible.

Less common SV references would still be generated as needed, but if even half of the references used were from the cache the generation time would be cut down proportionally. Thus requiring fewer resources on the least important step in the analysis pipeline.

READ CACHING

Each individually sequenced genome outputs billions of reads. In a non-tumor sequence 97% or more of these reads will align to the normal reference. At that point most analyses focus on the variations from the reference, which provides a far more compressed view of variations by simply counting the reads where a given variation is found. However, this is still an individual view of a genome. In order to compare at the population level analysis tools typically use information from one of the variation databases (e.g. dbSNP). The original aligned reads in the form of the BAM files are not discarded, however, they are not easily searchable either as they remain in individual sample files.

By taking advantage of distributed computing technologies this process could be reversed. Since a huge percentage of reads already align to the reference genome that means each BAM file has significant duplication both within the BAM and between individuals. Additionally, in order to compare two individuals it is necessary to directly analyze the variations separately before comparing them together. One method around this is to cache all of the reads in a data warehouse that is both scalable and searchable.

There are a variety of HPC database solutions available now that are specifically designed to scale up for “big data” (e.g. genomes, climate data, online digital photos, GPS information). Most of them use a flexible schema that enables the developer to dynamically modify the structure of stored data as it changes. Critically, all of them are distributed thus enabling the use of anything from a small cluster up to multiple data centers and cloud-based resources.

Using one of these distributed databases and cloud-based resources on the Amazon Elastic Compute Cloud a prototype read-caching system was tested. In this system sequence kmers (reads) are added to the cache. Each read is evaluated for the Levenshtein distance (LD) between it and all other reads with the same prefix kmer (1-4 bp), then added to the set of reads with the prefix. Metadata such as the aligned location, disease, and number of identical reads can be included. When provided with a new read or kmer sequence the caching system searches by prefix, and then the LD groups. This search method was found to be efficient regardless of data size: a search for a sequence kmer that existed in the cache took only 1.1 seconds, while searching for a unique read took 8.6 seconds to find the set of kmers that were most closely related by LD. Additionally, it was able to rapidly separate out unique reads from the common reads in two different individuals.

The primary drawback, and the reason this remains a prototype, is that populating the

cache initially involves a significant investment of time. This is consistent across all of the distributed databases (e.g. HBase, MongoDB, etc.), the initial set up requires significant investment. This means the cache should be live consistently requiring a cluster that can dedicate resources to keeping it available.

5.2.2 Data Integration

The focus on genomic information has greatly expanded our understanding of the complexity in tumor development and evolution. While a very small number of patients develop cancer due to heritable mutations most cases appear spontaneously with their own set of mutations. Mutational patterns have emerged that can connect carcinogens from tobacco and UV exposure to viral infections to tumors in specific tissues. Some structural mutations provide prognostic or therapeutic information about a patient's tumor as well. Genomic information in isolation cannot explain how all of those mutations may be accumulated, the huge diversity in tumor types, or predict a therapeutic response. This may be even more true in structural variation where most of the identified variants are of unknown origin, function, or importance. Answering these questions in genomic structural variation requires the integration of other types of data and various scales from epigenetics to help explain the origin of a mutation, to proteins that may explain the functional impact, and integrating with critical cancer pathways to understand the importance and possibly develop new therapies in response.

Epigenetics

An obvious integration point for chromosomal aberration data is epigenetic changes that could have specific effects on the genome. In addition to numerical and structural genomic instability cancer genomes are known to be epigenetically unstable. Hypermethylation is known to play a role in transcriptional silencing of tumor suppressor genes. Specific mutator phenotypes are also associated with epigenetic marks. For instance, the CpG island methylation phenotype (CIMP) is hypothesized to be involved with microsatellite instability in colorectal cancer, and has also been observed in glioblastomas, liver cancer, gastric cancer, ovarian cancer, and some leukemias (Issa 2004). However, the data on this is not entirely clear even within colorectal cancers. CIMP provides one phenotype that may be compared to other mutator phenotypes (e.g. CIN), but it needs to be connected to specific alterations beyond microsatellites. Methylation profiles of structurally aberrant genomes might help to support the connection. One area to target that is important in understanding the structural instability of tumor genomes is the regulation of DNA repair.

Methylation of specific histones is critical in the stabilization and activation

transcriptional elements involved in DSB repairs. These histones are regulated through methylation at the breakpoint, as well as phosphorylation of the appropriate transferase protein, though whether methylation is a dynamic process at the time of the break or a passive process due to a constant density is unclear. However, this methylation has to occur at the correct time to enable activation of the repair proteins. Ensuring the correct sequence of signaling and repair mechanisms occurs in response to DSBs is necessary to prevent further damage. Defects in the signaling events within the DNA damage checkpoints that negatively regulate damage response can result in incorrect repairs and structural instability.

Whether incorrect histone methylation is a cause or consequence of poor repair could provide further context on structural instability in cancer. It is also possible that this could provide the mechanism for the complex and sometimes sequential chromothriptic events seen in cancer genomes, as altered histones could be increasing the fragility of the individual cellular genome.

Pathway & Network Analysis

Recent transcriptional analysis of cancer cell lines suggests that identifying aberrant pathways is more effective when all of the mutation classes are integrated across all of the genes involved in the pathway (Klijn et al. 2014). While this analysis was a preliminary one, it suggested that looking at the mutational load in altered pathways and not the singular mutations could be of greater benefit. The same paper noted that certain genes, when involved in fusions, were more sensitive to targeted therapies regardless of their fusion partner. This suggests that a pathway analysis which weights the incidence of SVs that result in protein functional fusions could be useful in the search for targeted therapies.

As an example, in the TCGA data analyzed for this project a closer look at one of the breast cancer patients (see Table 11 in the Results chapter, patient BRCA (2)) shows that band 4q13 was highly represented in the tumor sample, but not in the germline. This was important for several reasons: 4q13 is known to integrate viral DNA from human papillomavirus (Kraus et al. 2008), which suggests that there are fragile sites for viral sequence insertion where structural variation could occur through other mechanisms as well. Furthermore, genes important to breast cancer development or aggressiveness are found in this region including EREG (and EPGN), which as a member of the epidermal growth factor gene family and structurally related to the ERBB tyrosine-kinase receptors, is involved in ER/HER2 status and tumor aggressiveness. In this thesis 17q23 was highly selected as one of the pairs for 4q13. While it has not been previously reported as having any fusions, it does harbor known breast cancer genes including BCAS3 (plays a role in angiogenesis) and BRCA1. It has also shown significant copy number gains in aggressive

breast cancers (Weber-Mangal et al. 2003).

If the region surrounding EREG were involved in the breakage/recombination it should be noted that the first order protein interactions that are involved are nearly all part of the protein tyrosine kinase signaling pathways based on gene ontology enrichment. Tyrosine kinases are critical for signaling pathways as well as frequently used therapeutic targets in various cancers. Additionally, several genes both up and downstream of EREG (BTC, AREG, EPGN) are involved in critical cancer pathways including EGFR and PI-3K. Both pathways are currently targeted using specific drug therapies. BCAS3 could be a more interesting problem. Multiple transcriptional variants are known for this gene across a 1Mb region of chromosome 17. The first order interactions with BCAS3 are mostly transcriptional regulation and known oncogenes (MTA1/2, TP53, BRCA1). Other genes in the surrounding region are also enriched for transcriptional regulation and DNA binding. Again, while the break may not directly involve these genes, it is likely that the pathways would be impacted. Interrogating the function of these pathways may provide evidence for the influence of a structural variation, or suggest a prognostic approach where sub-clonal tumor populations may make direct interrogation difficult. Ultimately, integrating genomic fusion information with pathways may enable the identification of sensitive drug targets, as well as network interactions that may provide for specific drug resistances within that pathway.

PROTEOMICS

The result of a structural variation in the genome is not necessarily a direct protein product as in the BCR-Abl example. While identifying aberrant proteins will continue to be useful for both their predictive and therapeutic values, there are other levels of proteomic data that could help to elucidate the consequence of breakpoints in the tumor genome. Regulation-specific phosphorylation sites have already been linked to altered interactions in critical signaling pathways. Metabolic interactions mediated by novel enzymes in a tumor have also been suggested as areas for biomarker discovery.

These novel enzymes could be identified, structurally, or functionally predicted based on known genomic alterations. Alternately, using qualitative mass spectrometry techniques novel proteins could be used to help identify the genomic alterations that produced them. Co-expression or altered gene regulation due to an aberration could also provide further information on biomarkers, again not directly related to the aberration but as an indirect consequence, which could provide new diagnostic methods in deep-tissue tumors (e.g. ovarian, prostate) that are generally difficult diagnose. This may be particularly relevant if expression alterations could alter post-translational modifications (e.g. phosphorylation) that are necessary to normal protein function.

It is necessary to keep in mind that proteomic effects are likely to be subtle and thus not directly predictive from a structural variation. While altered proteins could be translated from a variation region directly, it is far more likely that transcriptional alterations due to breakpoints in promoter or enhancer regions, or due to changes to the 3D structure of chromosomes may result in RNA products that interact with 'normal' proteins, or change the expression of a previously unrelated protein. Thus, while integrating proteomic information is going to be increasingly useful in identifying biomarkers, especially in critical pathways for tumorgenesis, it will be necessary to be aware of how complex the relationship is likely to be.

5.3 Conclusion

The holy grail of cancer research has long been a therapy that will eradicate all of the cancer cells in a patient. This has continued to prove elusive due to the complexity of the disease, heterogeneity within individual tumors, and the evolutionary processes that support progression and drug resistance. Large-scale structural variation is only one piece of this puzzle, but it has already provided crucial drug targets and biomarkers in patients. As with small-variants such as SNVs and indels, it is likely that many SV mutations will be neutral, or even deleterious. Knowing the full spectrum of survivable rearrangements that also promote tumor progression will enable the development of new interventions that are targeted to a specific patient's tumor, less toxic to the patient as a normal cell will not carry such rearrangements, unlikely to result in drug resistance as chromothripsis is such a chaotic event, and ultimately more effective.

Both experimental and technological issues have complicated elucidation of the extent and effect of chromosomal aberrations in the available sequencing data. Addressing some of these issues was the goal of this thesis. In conclusion the significant points of this thesis are that:

- *Structural variations are poorly understood*, as there is still a poor appreciation for the extent of structural variation in either the healthy or cancer genome. The MultiSieve method along with several new studies provides evidence that the extent and complexity of genomic rearrangement in the cancer genome is greater than current sequencing methods are capable of identifying. This also means that it is unclear just how many of these rearrangements may affect the progression of disease or may be of therapeutic value. While this thesis offers one method for unraveling this complexity it is likely that until long-read sequencing is appropriate for whole-genome usage structural variation will continue to complicate cancer analysis.

- *Variation identification is moving away from the reference.* In the search for both small (SNV) and large (SV) scale variations in cancer using one standard reference genome does not provide the necessary scope. Methods that avoid the reference entirely by directly comparing reads, or (like MultiSieve) that replace the single reference with hundreds of smaller references are necessary. This is because patient tumor samples are complex due to the heterogeneous mixture of cellular populations as well as the rearrangements from chromothriptic events. The MultiSieve method used all of the data available in a sequencing run as well, including reads that are unavailable to standard reference-based methods. This resulted in findings of greater complexity in tumor samples than were previously possible.
- *Scalable analyses are necessary to ask questions across populations.* This requires that current high-performance computing infrastructures are available and that appropriate software frameworks are adopted. There are now thousands of cancer genomes available to analyze, integrate, and otherwise include in new methods to identify biomarkers, driver mutations, or make predictions about subtypes. Using HPC frameworks and search algorithms developed specifically for sequence reads it becomes possible to both move away from the reference genome, and make cross population inferences and predictions regarding specific variations.
- *Karyotype to disease subtype mapping* aids in the process of variant identification in sequence read data. Some tumor subtypes and prognoses are already determined from the karyotype presentation. Using these karyotypes also provides direction and specificity to searching a space that encompasses 3 billion base pairs and tens of billions of sequence reads. It enables identification of individual regions of the genome to focus SV detection within, and evidence that at least one cellular population contains an aberration.
- *Using a knowledgebase of biological information directs the analysis of cancer genomes*. This is important due to the heterogeneity of the tumor (e.g. multiple cellular populations at varying sizes within the tumor) and the complexity of genomic aberrations that can occur as a result of a chromothriptic event. As the extent of small and large-scale structural variation in tumor genomes is still unclear the knowledgebase is incomplete, however using available knowledge (e.g. karyotype aberrations currently) both pan-cancer and tumor-specific aberration information can inform the search for novel and known aberrations in a patient's tumor. Improving this knowledgebase with identifications of structural variants will

continue to enable a more directed search and will continue to provide information to direct targeted long-read sequencing to verify or identify SVs.

- *Coverage differences within a mixed population sample must be taken into account.* MultiSieve highlighted the complexity of genomic variations, but it also showed that coverage differences across the different populations within a sample are important to the identification of SVs in smaller sub-clonal populations. Using the positional clustering parameter in the score methodology it is possible to select for regions that are represented by small sub-populations and are therefore uncommon in the sequence reads.
- *Tumor heterogeneity and sub-clonal mutations are detectable* when using multiple references. By adding reads from a breakpoint at known coverage to patient samples it was shown that sub-clonal cellular populations can be identified using this method. Combining this information with small-scale variant mutational pattern identification for sub-clonal populations it may be possible to predict disease progression, for instance if a SV is an early mutation it may be indicative of later disease behavior. The MultiSieve method provides a necessary first step to connecting these two different analyses in cancer variation.

This thesis provided an innovative method to approach these issues through the use of available knowledge, high performance computing, and changing the assumptions behind the model of a cancer genome. It will continue to be a viable approach when long-reads become a standard technique for sequencing, as longer reads still require localization in the known genomic space. Finally, this approach will improve in computational efficiency and accuracy as new variants are identified and included in the background knowledge further enabling targeted and rapid identification for already known structural variants and supporting the search for new ones.

REFERENCES

Abo, Ryan P, Matthew Ducar, Elizabeth P Garcia, Aaron R Thorner, Vanesa Rojas-Rudilla, Ling Lin, Lynette M Sholl, et al. 2014. “BreaKmer: Detection of Structural Variation in Targeted Massively Parallel Sequencing Data Using Kmers.” *Nucleic Acids Research* 43 (3) (November 26): e19. doi:10.1093/nar/gku1211.

Aho, Alfred V., and Margaret J. Corasick. 1975. “Efficient String Matching: An Aid to Bibliographic Search.” *Communications of the ACM* 18 (6) (June 1): 333–340. doi:10.1145/360825.360855.

Aird, Daniel, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. “Analyzing and Minimizing PCR Amplification Bias in Illumina Sequencing Libraries.” *Genome Biology* 12 (2) (January): R18. doi:10.1186/gb-2011-12-2-r18.

Al-Hajj, Muhammad, Max S Wicha, Adalberto Benito-Hernandez, Sean J Morrison, and Michael F Clarke. 2003. “Prospective Identification of Tumorigenic Breast Cancer Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (7) (April 1): 3983–8. doi:10.1073/pnas.0530291100.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel a. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* (August 14). doi:10.1038/nature12477.

Attolini, Camille Stephan-Otto, Yu-Kang Cheng, Rameen Beroukhim, Gad Getz, Omar Abdel-Wahab, Ross L Levine, Ingo K Mellinghoff, and Franziska Michor. 2010. “A Mathematical Framework to Determine the Temporal Sequence of Somatic Genetic Events in Cancer.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (41) (October 12): 17604–9. doi:10.1073/pnas.1009117107.

Baca, Sylvan C, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, et al. 2013. “Punctuated Evolution of Prostate Cancer Genomes.” *Cell* 153 (3) (April 25): 666–77. doi:10.1016/j.cell.2013.03.021.

Baker, Monya. 2012. “De Novo Genome Assembly: What Every Biologist Should Know.” *Nature Methods* 9 (4) (March 27): 333–337. doi:10.1038/nmeth.1935.

Bamford, S, E Dawson, S Forbes, J Clements, R Pettett, a Dogan, a Flanagan, et al. 2004. “The COSMIC (Catalogue of Somatic Mutations in Cancer) Database and Website.” *British Journal of Cancer* 91 (2) (July 19): 355–8. doi:10.1038/sj.bjc.6601894.

Barenboim-Stapleton, Linda, Xuezhong Yang, Maria Tsokos, Jon M Wigginton, Hesed Padilla-Nash, Thomas Ried, and Carol J Thiele. 2005. “Pediatric Pancreatoblastoma: Histopathologic and Cytogenetic Characterization of Tumor and Derived Cell Line.”

Cancer Genetics and Cytogenetics 157 (2) (March): 109–17. doi:10.1016/j.cancergencyto.2004.05.017.

Barlow, Jacqueline H, Robert B Faryabi, Elsa Callén, Nancy Wong, Amy Malhowski, Hua Tang Chen, Gustavo Gutierrez-Cruz, et al. 2013. "Identification of Early Replicating Fragile Sites That Contribute to Genome Instability." *Cell* 152 (3) (January 31): 620–32. doi:10.1016/j.cell.2013.01.006.

Benjamini, Yuval, and Terence P. Speed. 2012. "Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing." *Nucleic Acids Research* 40 (10): 1–14. doi:10.1093/nar/gks001.

Bennett, Charles H, Péter Gács, Ming Li, Paul M B Vitányi, and Wojciech H Zurek. 1998. "Information Distance." *Information Theory, IEEE Transactions on* 44 (4): 1407–1423.

Berger, Michael F, Michael S Lawrence, Francesca Demichelis, Yotam Drier, Kristian Cibulskis, Andrey Y Sivachenko, Andrea Sboner, et al. 2011. "The Genomic Complexity of Primary Human Prostate Cancer." *Nature* 470 (7333) (February 10): 214–20.

Bignell, Graham R, Chris D Greenman, Helen Davies, Adam P Butler, Sarah Edkins, Jenny M Andrews, Gemma Buck, et al. 2010. "Signatures of Mutation and Selection in the Cancer Genome." *Nature* 463 (7283) (February 18): 893–8. doi:10.1038/nature08768.

Bilbao, Cristina, Raquel Ramírez, Germán Rodríguez, Orlando Falcón, Laureano León, Nicolás Díaz-Chico, Manuel Perucho, and Juan Carlos Díaz-Chico. 2010. "Double Strand Break Repair Components Are Frequent Targets of Microsatellite Instability in Endometrial Cancer." *European Journal of Cancer (Oxford, England : 1990)* 46 (15) (October): 2821–7. doi:10.1016/j.ejca.2010.06.116.

Bolli, Niccolo, Hervé Avet-Loiseau, David C Wedge, Peter Van Loo, Ludmil B Alexandrov, Inigo Martincorena, Kevin J Dawson, et al. 2014. "Heterogeneity of Genomic Evolution and Mutational Profiles in Multiple Myeloma." *Nature Communications* 5 (January): 2997. doi:10.1038/ncomms3997.

Bonnet, Dominique, and John E. Dick. 1997. "Human Acute Myeloid Leukemia Is Organized as a Hierarchy That Originates from a Primitive Hematopoietic Cell." *Nature Medicine* 3 (7) (July): 730–737. doi:10.1038/nm0797-730.

Bracken, Adrian P, Diego Pasini, Maria Capra, Elena Prosperini, Elena Colli, and Kristian Helin. 2003. "EZH2 Is Downstream of the pRB-E2F Pathway, Essential for Proliferation and Amplified in Cancer." *The EMBO Journal* 22 (20) (October 15): 5323–35. doi:10.1093/emboj/cdg542.

Burrack, Laura S, and Judith Berman. 2012. "Flexibility of Centromere and Kinetochore Structures." *Trends in Genetics : TIG* 28 (5) (May): 204–12. doi:10.1016/j.tig.2012.02.003.

Burrell, Rebecca A., Sarah E. McClelland, David Endesfelder, Petra Groth, Marie-Christine Weller, Nadeem Shaikh, Enric Domingo, et al. 2013. “Replication Stress Links Structural and Numerical Cancer Chromosomal Instability.” *Nature* 494 (7438) (February 27): 492–496. doi:10.1038/nature11935.

Burrell, Rebecca a., Nicholas McGranahan, Jiri Bartek, and Charles Swanton. 2013. “The Causes and Consequences of Genetic Heterogeneity in Cancer Evolution.” *Nature* 501 (7467) (September 18): 338–345. doi:10.1038/nature12625.

Campbell, Peter J, Philip J Stephens, Erin D Pleasance, Sarah O’Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, et al. 2008. “Identification of Somatically Acquired Rearrangements in Cancer Using Genome-Wide Massively Parallel Paired-End Sequencing.” *Nature Genetics* 40 (6) (June): 722–9. doi:10.1038/ng.128.

Carneiro, Mauricio O, Carsten Russ, Michael G Ross, Stacey B Gabriel, Chad Nusbaum, and Mark A DePristo. 2012. “Pacific Biosciences Sequencing Technology for Genotyping and Variation Discovery in Human Data.” *BMC Genomics* 13 (January): 375. doi:10.1186/1471-2164-13-375.

Chen, Ken, John W Wallis, Michael D Mclellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D Mcgrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and R Elaine. 2009. “BreakDancer: An Algorithm for High Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6 (9): 677–681. doi:10.1038/nmeth.1363.BreakDancer.

Chen, Ken, John W Wallis, Michael D Mclellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D Mcgrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, Elaine R Mardis, et al. 2009. “BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation.” *Nature Methods* 6 (9) (September): 677–81. doi:10.1038/nmeth.1363.BreakDancer.

Chen, Yuan, Fiona Cunningham, Daniel Rios, William M McLaren, James Smith, Bethan Pritchard, Giulietta M Spudich, et al. 2010. “Ensembl Variation Resources.” *BMC Genomics* 11 (1) (January): 293. doi:10.1186/1471-2164-11-293.

Cheng, Yu-Wei, Hanna Pincas, Manny D Bacolod, Gunter Schemmann, Sarah F Giardina, Jianmin Huang, Sandra Barral, et al. 2008. “CpG Island Methylator Phenotype Associates with Low-Degree Chromosomal Abnormalities in Colorectal Cancer.” *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 14 (19) (October 1): 6005–13. doi:10.1158/1078-0432.CCR-08-0216.

Chris Fraley, Adrian E. Raftery. 2002. “Model-Based Clustering, Discriminant Analysis, and Density Estimation.” *Journal of the American Statistical Association* 97 (458): 611–631.

Cilibrasi, Rudi, and Paul M B Vitányi. 2005. “Clustering by Compression.” *IEEE Transactions on Information Theory* 51 (4): 1523–1545.

Corthay, Alexandre, Dag K Skovseth, Katrin U Lundin, Egil Røsjø, Hilde Omholt, Peter O Hofgaard, Guttorm Haraldsen, and Bjarne Bogen. 2005. “Primary Antitumor Immune Response Mediated by CD4+ T Cells.” *Immunity* 22 (3) (March): 371–83. doi:10.1016/j.immuni.2005.02.003.

Damerau, Fred J. 1964. “A Technique for Computer Detection and Correction of Spelling Errors.” *Communications of the ACM* 7 (3) (March 1): 171–176. doi:10.1145/363958.363994.

Das, Rajdeep, Nevenka Dimitrova, Zhenyu Xuan, Robert A Rollins, Fatemah Haghighi, John R Edwards, Jingyue Ju, Timothy H Bestor, and Michael Q Zhang. 2006. “Computational Prediction of Methylation Status in Human Genomic Sequences.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (28) (July 11): 10713–6. doi:10.1073/pnas.0602949103.

Dean, Jeffrey, and Sanjay Ghemawat. 2008. “MapReduce : Simplified Data Processing on Large Clusters.” Edited by L Purich Daniel. *Communications of the ACM* 51 (1). SIGMOD ’07: 1–13. doi:10.1145/1327452.1327492.

Derrien, Thomas, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. 2012. “Fast Computation and Applications of Genome Mappability.” *PloS One* 7 (1) (January): e30377. doi:10.1371/journal.pone.0030377.

Ding, Li, Matthew J Ellis, Shunqiang Li, David E Larson, Ken Chen, John W Wallis, Christopher C Harris, et al. 2010. “Genome Remodelling in a Basal-like Breast Cancer Metastasis and Xenograft.” *Nature* 464 (7291) (April 15): 999–1005. doi:10.1038/nature08989.

Ding, Li, Michael C Wendl, Daniel C Koboldt, and Elaine R Mardis. 2010. “Analysis of next-Generation Genomic Data in Cancer: Accomplishments and Challenges.” *Human Molecular Genetics* 19 (R2) (October 15): R188–96. doi:10.1093/hmg/ddq391.

Duijf, P H G, and R Benezra. 2013. “The Cancer Biology of Whole-Chromosome Instability.” *Oncogene* 32 (40) (January 14): 4727–4736. doi:10.1038/onc.2012.616.

Dunican, Donncha S, Peter McWilliam, Orna Tighe, Anne Parle-McDermott, and David T Croke. 2002. “Gene Expression Differences between the Microsatellite Instability (MIN) and Chromosomal Instability (CIN) Phenotypes in Colorectal Cancer Revealed by High-Density cDNA Array Hybridization.” *Oncogene* 21 (20) (May 9): 3253–7. doi:10.1038/sj.onc.1205431.

Edwards, Paul. 2012. “University of Cambridge SKY Karyotypes and FISH Analysis of Epithelial Cancer Cell Lines.” Accessed October 22. http://www.path.cam.ac.uk/~pawefish/.

Eltarhouny, S A, W H Elsawy, R Radpour, S Hahn, W Holzgreve, and X Y Zhong. 2008. “Genes Controlling Spread of Breast Cancer to Lung ‘Gang of 4’.” *Experimental Oncology* 30 (2) (June): 91–5.

Fearon, E. R., P. J. Burke, C. A. Schiffer, B. A. Zehnbauer, and B. Vogelstein. 1986. "Differentiation of Leukemia Cells to Polymorphonuclear Leukocytes in Patients with Acute Nonlymphocytic Leukemia." *New England Journal of Medicine* 315 (1): 15–24.

FitzPatrick, David R, Jacqueline Ramsay, Niolette I McGill, Mary Shade, Andrew D Carothers, and Nicholas D Hastie. 2002. "Transcriptome Analysis of Human Autosomal Trisomy." *Human Molecular Genetics* 11 (26) (December 15): 3249–56.

Gajewski, Thomas F, Hans Schreiber, and Yang-Xin Fu. 2013. "Innate and Adaptive Immune Cells in the Tumor Microenvironment." *Nature Immunology* 14 (10) (October): 1014–22. doi:10.1038/ni.2703.

Gardner, Malcolm J, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, et al. 2002. "Genome Sequence of the Human Malaria Parasite Plasmodium Falciparum." *Nature* 419 (6906) (October 3): 498–511. doi:10.1038/nature01097.

Gill, Steven R, Mihai Pop, Robert T Deboy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. 2006. "Metagenomic Analysis of the Human Distal Gut Microbiome." *Science (New York, N.Y.)* 312 (5778) (June 2): 1355–9. doi:10.1126/science.1124234.

Greaves, Mel, and Carlo C Maley. 2012. "Clonal Evolution in Cancer." *Nature* 481 (7381) (January 19): 306–13. doi:10.1038/nature10762.

Guerrero, Astrid Alonso, Mercedes Cano Gamero, Varvara Trachana, Agnes Fütterer, Cristina Pacios-Bras, Nuria Panadero Díaz-Concha, Juan Cruz Cigudosa, Carlos Martínez-A, and Karel H M van Wely. 2010. "Centromere-Localized Breaks Indicate the Generation of DNA Damage by the Mitotic Spindle." *Proceedings of the National Academy of Sciences of the United States of America* 107 (9) (March 2): 4159–64. doi:10.1073/pnas.0912143106.

Hach, Faraz, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E Eichler, and S Cenk Sahinalp. 2010. "mrsFAST: A Cache-Oblivious Algorithm for Short-Read Mapping." *Nature Methods* 7 (8) (August): 576–7. doi:10.1038/nmeth0810-576.

Hamosh, Ada, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. 2005. "Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders." *Nucleic Acids Research* 33 (Database issue) (January 1): D514–7. doi:10.1093/nar/gki033.

Hanahan, Douglas, and Robert A Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144 (5) (March 4): 646–74. doi:10.1016/j.cell.2011.02.013.

Hart, Steven N, Vivekananda Sarangi, Raymond Moore, Saurabh Baheti, Jaysheel D Bhavsar, Fergus J Couch, and Jean-Pierre a Kocher. 2013. "SoftSearch: Integration of Multiple Sequence Features to Identify Breakpoints of Structural Variations." *PloS One* 8 (12) (January): e83356. doi:10.1371/journal.pone.0083356.

Hayflick, L. 2000. “The Illusion of Cell Immortality.” *British Journal of Cancer* 83 (7) (October): 841–6. doi:10.1054/bjoc.2000.1296.

Hickman, E. 2002. “The Role of p53 and pRB in Apoptosis and Cancer.” *Current Opinion in Genetics & Development* 12 (1) (February 1): 60–66. doi:10.1016/S0959-437X(01)00265-9.

Holstege, Henne, Wayne Pfeiffer, Daoud Sie, Marc Hulsman, Thomas J Nicholas, Clarence C Lee, Tristen Ross, et al. 2014. “Somatic Mutations Found in the Healthy Blood Compartment of a 115-Yr-Old Woman Demonstrate Oligoclonal Hematopoiesis.” *Genome Research* 24 (5) (April 23): 733–742. doi:10.1101/gr.162131.113.

Hormozdiari, Fereydoun, Iman Hajirasouliha, Phuong Dao, Faraz Hach, Deniz Yorukoglu, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. 2010. “Next-Generation VariationHunter: Combinatorial Algorithms for Transposon Insertion Discovery.” *Bioinformatics (Oxford, England)* 26 (12) (June 15): i350–7.

Hormozdiari, Fereydoun, Iman Hajirasouliha, Andrew McPherson, Evan E. Eichler, and S. Cenk Sahinalp. 2011. “Simultaneous Structural Variation Discovery among Multiple Paired-End Sequenced Genomes.” *Genome Research* 21 (12): 2203–2212. doi:10.1101/gr.120501.111.

Hou, Yong, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, et al. 2012. “Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm.” *Cell* 148 (5) (March 2): 873–85. doi:10.1016/j.cell.2012.02.028.

Huang, Weichun, Leping Li, Jason R Myers, and Gabor T Marth. 2012. “ART: A next-Generation Sequencing Read Simulator.” *Bioinformatics (Oxford, England)* 28 (4) (February 15): 593–4. doi:10.1093/bioinformatics/btr708.

IJdo, J W, a Baldini, D C Ward, S T Reeders, and R a Wells. 1991. “Origin of Human Chromosome 2: An Ancestral Telomere-Telomere Fusion.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (20): 9051–9055. doi:10.1073/pnas.88.20.9051.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011) (October 21): 931–45. doi:10.1038/nature03001.

Issa, Jean-Pierre. 2004. “CpG Island Methylator Phenotype in Cancer.” *Nature Reviews. Cancer* 4 (12) (December): 988–93. doi:10.1038/nrc1507.

Janssen, A, and R H Medema. 2012. “Genetic Instability: Tipping the Balance.” *Oncogene* 32 (38) (December 17): 4459–4470. doi:10.1038/onc.2012.576.

Jiang, Yue, Yadong Wang, and Michael Brudno. 2012. “PRISM: Pair-Read Informed Split-Read Mapping for Base-Pair Level Detection of Insertion, Deletion and Structural

Variants." *Bioinformatics (Oxford, England)* 28 (20) (October 15): 2576–83. doi:10.1093/bioinformatics/bts484.

Jiao, Wei, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. 2014. "Inferring Clonal Evolution of Tumors from Single Nucleotide Somatic Mutations." *BMC Bioinformatics* 15 (1) (January): 35. doi:10.1186/1471-2105-15-35.

Kaelin, William G. 2004. "Gleevec: Prototype or Outlier?" *Science's STKE : Signal Transduction Knowledge Environment* 2004 (225) (March 23): pe12. doi:10.1126/stke.2252004pe12.

Katzman, Sol, John A Capra, David Haussler, and Katherine S Pollard. 2011. "Ongoing GC-Biased Evolution Is Widespread in the Human Genome and Enriched near Recombination Hot Spots." *Genome Biology and Evolution* 3 (January): 614–26. doi:10.1093/gbe/evr058.

Killcoyne, Sarah, and Antonio del Sol. 2014. "FIGG: Simulating Populations of Whole Genome Sequences for Heterogeneous Data Analyses." *BMC Bioinformatics* 15 (1): 149. doi:10.1186/1471-2105-15-149.

Killcoyne, Sarah, and Antonio del Sol. 2015. "Identification of Large-Scale Genomic Variation in Cancer Genomes Using in Silico Reference Models." *Nucleic Acids Research* (August 11): gkv828. doi:10.1093/nar/gkv828.

Kim, M S, S S Kim, E M Je, N J Yoo, and S H Lee. 2012. "Mutational and Expressional Analyses of STAG2 Gene in Solid Cancers." *Neoplasma* 59 (5) (January): 524–9. doi:10.4149/neo_2012_067.

Kirsch, I R, E D Green, R Yonescu, R Strausberg, N Carter, D Bentley, M A Leversha, et al. 2000. "A Systematic, High-Resolution Linkage of the Cytogenetic and Physical Maps of the Human Genome." *Nature Genetics* 24 (4) (April): 339–40. doi:10.1038/74149.

Kitts, Adrienne, and Stephen Sherry. 2002. *The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation*. National Center for Biotechnology Information (US).

Klijn, Christiaan, Steffen Durinck, Eric W Stawiski, Peter M Haverty, Zhaoshi Jiang, Hanbin Liu, Jeremiah Degenhardt, et al. 2014. "A Comprehensive Transcriptional Portrait of Human Cancer Cell Lines." *Nature Biotechnology* 33 (3) (December 8): 306–312. doi:10.1038/nbt.3080.

Kraus, Irene, Corina Driesch, Svetlana Vinokurova, Eivind Hovig, Achim Schneider, Magnus von Knebel Doeberitz, and Matthias Dürst. 2008. "The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes." *Cancer Research* 68 (7) (April 1): 2514–22. doi:10.1158/0008-5472.CAN-07-2776.

Kudla, Grzegorz, Aleksandra Helwak, and Leszek Lipinski. 2004. “Gene Conversion and GC-Content Evolution in Mammalian Hsp70.” *Molecular Biology and Evolution* 21 (7) (July 1): 1438–44. doi:10.1093/molbev/msh146.

Lander, E S, and M S Waterman. 1988. “Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis.” *Genomics* 2 (3) (April): 231–9.

Langmead, Ben, Michael C Schatz, Jimmy Lin, Mihai Pop, and Steven L Salzberg. 2009. “Searching for SNPs with Cloud Computing.” *Genome Biology* 10 (11) (January): R134. doi:10.1186/gb-2009-10-11-r134.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3) (January): R25. doi:10.1186/gb-2009-10-3-r25.

Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* (June 16): 1–5. doi:10.1038/nature12213.

Lee, Seunghak, Fereydoun Hormozdiari, Can Alkan, and Michael Brudno. 2009. “MoDIL: Detecting Small Indels from Clone-End Sequencing with Mixtures of Distributions.” *Nature Methods* 6 (7): 473–474.

Lercher, Martin J, and Laurence D Hurst. 2002. “Human SNP Variability and Mutation Rate Are Higher in Regions of High Recombination.” *Trends in Genetics* 18 (7) (July): 337–340. doi:10.1016/S0168-9525(02)02669-0.

Levenshtein, Vladimir I. 1966. “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.” *Soviet Physics Doklady*.

Lewis, Steven, Attila Csordas, Sarah Killcoyne, Henning Hermjakob, Michael Hoopmann, Robert Moritz, Eric Deutsch, and John Boyle. 2012. “Hydra: A Scalable Proteomic Search Engine Which Utilizes the Hadoop Distributed Computing Framework.” *BMC Bioinformatics* 13 (1): 324.

Lewis, Steven, Sheila Reynolds, Hector Rovera, Mike O’Leary, Sarah Killcoyne, Ilya Shmulevich, and John Boyle. 2010. “Howdah - A Flexible Pipeline Framework for Analyzing Genomic Data.” In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, 776–779. IEEE. doi:10.1109/CloudCom.2010.75.

Ley, Timothy J, Elaine R Mardis, Li Ding, Bob Fulton, Michael D Mclellan, Ken Chen, David Dooling, et al. 2008. “DNA Sequencing of a Cytogenetically Normal Acute Myeloid Leukemia Genome.” *Nature* 456 (7218): 66–72. doi:10.1038/nature07485.DNA.

Li, H.-R. 2004. “Hypersensitivity of Tumor Cell Lines with Microsatellite Instability to DNA Double Strand Break Producing Chemotherapeutic Agent Bleomycin.” *Cancer Research* 64 (14) (July 15): 4760–4767. doi:10.1158/0008-5472.CAN-04-0975.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14) (July 15): 1754–60. doi:10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16) (August 15): 2078–9. doi:10.1093/bioinformatics/btp352.

Li, Heng, and Nils Homer. 2010. “A Survey of Sequence Alignment Algorithms for next-Generation Sequencing.” *Briefings in Bioinformatics* 11 (5) (September 1): 473–83. doi:10.1093/bib/bbq015.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11) (November 1): 1851–8. doi:10.1101/gr.078212.108.

Li, Runsheng, Chia-Ling Hsieh, Amanda Young, Zhihong Zhang, Xiaoliang Ren, and Zhongying Zhao. 2015. “Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences Even in the ‘Finished’ C. Elegans Genome.” *Scientific Reports* 5 (January 3): 10814. doi:10.1038/srep10814.

Loeb, Lawrence A. 2001. “A Mutator Phenotype in Cancer.” *Cancer Res.* 61 (8) (April 1): 3230–3239.

Look, A. T. 1997. “Oncogenic Transcription Factors in the Human Acute Leukemias.” *Science* 278 (5340) (November 7): 1059–1064. doi:10.1126/science.278.5340.1059.

Lynch, Michael. 2010. “Rate, Molecular Spectrum, and Consequences of Human Mutation.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (3) (January 19): 961–8. doi:10.1073/pnas.0912629107.

Manning, a L, C Benes, and N J Dyson. 2013. “Whole Chromosome Instability Resulting from the Synergistic Effects of pRB and p53 Inactivation.” *Oncogene* (April) (June 24): 1–8. doi:10.1038/onc.2013.201.

McBride, David J, Dariush Etemadmoghadam, Susanna L Cooke, Kathryn Alsop, Joshy George, Adam Butler, Juok Cho, et al. 2012. “Tandem Duplication of Chromosomal Segments Is Common in Ovarian and Breast Cancer Genomes.” *The Journal of Pathology* 227 (4) (August): 446–55. doi:10.1002/path.4042.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9) (September 1): 1297–303. doi:10.1101/gr.107524.110.

Meacham, Corbin E., and Sean J. Morrison. 2013. “Tumour Heterogeneity and Cancer Cell Plasticity.” *Nature* 501 (7467) (September 18): 328–337. doi:10.1038/nature12624.

Medvedev, Paul, Monica Stanciu, and Michael Brudno. 2009. “Computational Methods for Discovering Structural Variation with next-Generation Sequencing.” *Nature Methods* 6 (11 Suppl) (November): S13–20. doi:10.1038/nmeth.1374.

Mitelman, F, B Johansson, and F (Eds.) Mertens. 2015. “Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.” http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Mitelman, Felix, Bertil Johansson, and Fredrik Mertens. 2007. “The Impact of Translocations and Gene Fusions on Cancer Causation.” *Nature Reviews. Cancer* 7 (4) (April): 233–45. doi:10.1038/nrc2091.

Moncunill, Valentí, Santi Gonzalez, Sílvia Beà, Lise O Andrieux, Itziar Salaverria, Cristina Royo, Laura Martinez, et al. 2014. “Comprehensive Characterization of Complex Structural Variations in Cancer by Directly Comparing Genome Sequence Reads.” *Nature Biotechnology* (October 26). doi:10.1038/nbt.3027.

Moynahan, Mary Ellen, and Maria Jasin. 2010. “Mitotic Homologous Recombination Maintains Genomic Stability and Suppresses Tumorigenesis.” *Nature Reviews. Molecular Cell Biology* 11 (3) (March): 196–207. doi:10.1038/nrm2851.

Nambiar, Mridula, Vijayalakshmi Kari, and Sathees C Raghavan. 2008. “Chromosomal Translocations in Cancer.” *Biochimica et Biophysica Acta* 1786 (2) (December): 139–52. doi:10.1016/j.bbcan.2008.07.005.

“NCI and NCBI’s SKY/M-FISH and CGH Database.” 2012. Accessed November 12. http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi.

“NCI Fredrick National Laboratory Cell Line Drug Discovery Panel.” 2012. Accessed November 26. http://home.ncifcrf.gov/CCR/60SKY/new/demo1.asp.

Niemenmaa, Matti, Aleksi Kallio, André Schumacher, Petri Klemelä, Eija Korpelainen, and Keijo Heljanko. 2012. “Hadoop-BAM: Directly Manipulating next Generation Sequencing Data in the Cloud.” *Bioinformatics (Oxford, England)* 28 (6) (March 15): 876–7. doi:10.1093/bioinformatics/bts054.

Nik-Zainal, Serena, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell* 149 (5) (May 25): 994–1007. doi:10.1016/j.cell.2012.04.023.

Nishida, Naoyo, Hirohisa Yano, Takashi Nishida, Toshiharu Kamura, and Masamichi Kojiro. 2006. “Angiogenesis in Cancer.” *Vascular Health and Risk Management* 2 (3) (January): 213–9.

Nowell, P. 1976. “The Clonal Evolution of Tumor Cell Populations.” *Science* 194 (4260) (October 1): 23–28. doi:10.1126/science.959840.

Nowell, P., and D. Hungerford. 1961. “Chromosome Studies in Human Leukemia. II. Chronic Granulocytic Leukemia.” *Journal of the National Cancer Institute* 27 (November): 1013–35.

O'Brien, Catherine A, Aaron Pollett, Steven Gallinger, and John E Dick. 2007. "A Human Colon Cancer Cell Capable of Initiating Tumour Growth in Immunodeficient Mice." *Nature* 445 (7123) (January 4): 106–10. doi:10.1038/nature05372.

Pinard, Robert, Alex de Winter, Gary J Sarkis, Mark B Gerstein, Karrie R Tartaro, Ramona N Plant, Michael Egholm, Jonathan M Rothberg, and John H Leamon. 2006. "Assessment of Whole Genome Amplification-Induced Bias through High-Throughput, Massively Parallel Whole Genome Sequencing." *BMC Genomics* 7 (1) (January): 216. doi:10.1186/1471-2164-7-216.

Pleasance, Erin D, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, et al. 2010. "A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome." *Nature* 463 (7278) (January 14): 191–6. doi:10.1038/nature08658.

Pleasance, Erin D, Philip J Stephens, Sarah O'Meara, David J McBride, Alison Meynert, David Jones, Meng-Lay Lin, et al. 2010. "A Small-Cell Lung Cancer Genome with Complex Signatures of Tobacco Exposure." *Nature* 463 (7278) (January 14): 184–90. doi:10.1038/nature08629.

Przybytkowski, Ewa, Elizabeth Lenkiewicz, Michael T Barrett, Kathleen Klein, Sheida Nabavi, Celia Mt Greenwood, and Mark Basik. 2014. "Chromosome-Breakage Genomic Instability and Chromothripsis in Breast Cancer." *BMC Genomics* 15 (1): 579. doi:10.1186/1471-2164-15-579.

Rambaut, Andrew, Oliver G. Pybus, Martha I. Nelson, Cecile Viboud, Jeffery K. Taubenberger, and Edward C. Holmes. 2008. "The Genomic and Epidemiological Dynamics of Human Influenza A Virus." *Nature* 453 (7195) (April 16): 615–619. doi:10.1038/nature06945.

Rausch, T., T. Zichner, a. Schlattl, a. M. Stutz, V. Benes, and J. O. Korbel. 2012. "DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis." *Bioinformatics* 28 (18) (September 7): i333–i339. doi:10.1093/bioinformatics/bts378.

Ried, K. 2000. "Common Chromosomal Fragile Site FRA16D Sequence: Identification of the FOR Gene Spanning FRA16D and Homozygous Deletions and Translocation Breakpoints in Cancer Cells." *Human Molecular Genetics* 9 (11) (July 1): 1651–1663. doi:10.1093/hmg/9.11.1651.

Robinson, Dan R, Shanker Kalyana-Sundaram, Yi-Mi Wu, Sunita Shankar, Xuhong Cao, Bushra Ateeq, Irfan A Asangani, et al. 2011. "Functionally Recurrent Rearrangements of the MAST Kinase and Notch Gene Families in Breast Cancer." *Nature Medicine* 17 (12) (December): 1646–51. doi:10.1038/nm.2580.

Robinson, Thomas, Sarah Killcoyne, Ryan Bressler, and John Boyle. 2011. "SAMQA: Error Classification and Validation of High-Throughput Sequenced Read Data." *BMC Genomics* 12 (1): 419. doi:10.1186/1471-2164-12-419.

Ruffalo, Matthew, Thomas LaFramboise, and Mehmet Koyutürk. 2011. "Comparative Analysis of Algorithms for next-Generation Sequencing Read Alignment."

Bioinformatics (Oxford, England) 27 (20) (October 15): 2790–6. doi:10.1093/bioinformatics/btr477.

Schatz, Michael C. 2009. “CloudBurst: Highly Sensitive Read Mapping with MapReduce.” *Bioinformatics (Oxford, England)* 25 (11) (June 1): 1363–9. doi:10.1093/bioinformatics/btp236.

Schatz, Michael C, Arthur L Delcher, and Steven L Salzberg. 2010. “Assembly of Large Genomes Using Second-Generation Sequencing.” *Genome Research* 20 (9) (September): 1165–73. doi:10.1101/gr.101360.109.

Schatz, Michael C, Ben Langmead, and Steven L Salzberg. 2010. “Cloud Computing and the DNA Data Race.” *Nature Biotechnology* 28 (7) (July): 691–3. doi:10.1038/nbt0710-691.

Schbath, Sophie, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, and Jean-François Gibrat. 2012. “Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis.” *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 19 (6) (June): 796–813. doi:10.1089/cmb.2012.0022.

Schröck, E, S du Manoir, T Veldman, B Schoell, J Wienberg, M A Ferguson-Smith, Y Ning, et al. 1996. “Multicolor Spectral Karyotyping of Human Chromosomes.” *Science (New York, N.Y.)* 273 (5274) (July 26): 494–7.

Shackleton, Mark, Elsa Quintana, Eric R. Fearon, and Sean J. Morrison. 2009. “Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution.” *Cell* 138 (5): 822–829. doi:10.1016/j.cell.2009.08.017.

Shah, Nameeta, Michael Lankerovich, Hwahyung Lee, Jae-Geun Yoon, Brett Schroeder, and Greg Foltz. 2013. “Exploration of the Gene Fusion Landscape of Glioblastoma Using Transcriptome Sequencing and Copy Number Data.” *BMC Genomics* 14 (1): 818. doi:10.1186/1471-2164-14-818.

Shigesada, Katsuya, Bart van de Sluis, and P Paul Liu. 2004. “Mechanism of Leukemogenesis by the inv(16) Chimeric Gene CBFB/PEBP2B-MHY11.” *Oncogene* 23 (24) (May 24): 4297–307. doi:10.1038/sj.onc.1207748.

Singh, Sheila K, Cynthia Hawkins, Ian D Clarke, Jeremy A Squire, Jane Bayani, Takuichiro Hide, R Mark Henkelman, Michael D Cusimano, and Peter B Dirks. 2004. “Identification of Human Brain Tumour Initiating Cells.” *Nature* 432 (7015) (November 18): 396–401. doi:10.1038/nature03128.

Speicher, M R, S Gwyn Ballard, and D C Ward. 1996. “Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH.” *Nature Genetics* 12 (4) (April): 368–75. doi:10.1038/ng0496-368.

Stephens, Philip J, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, et al. 2011. “Massive Genomic Rearrangement

Acquired in a Single Catastrophic Event during Cancer Development." *Cell* 144 (1) (January 7): 27–40.

Stewart, Jocelyn M, Patricia A Shaw, Craig Gedye, Marcus Q Bernardini, Benjamin G Neel, and Laurie E Ailles. 2011. "Phenotypic Heterogeneity and Instability of Human Ovarian Tumor-Initiating Cells." *Proceedings of the National Academy of Sciences of the United States of America* 108 (16) (April 19): 6468–73. doi:10.1073/pnas.1005529108.

Storn, Rainer, and Kenneth Price. 1997. "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces." *Journal of Global Optimization* 11: 341–359.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239) (April 9): 719–24. doi:10.1038/nature07943.

The 1000 Genomes Project Consortium. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319) (October 28): 1061–73. doi:10.1038/nature09534.

The Apache Software Foundation. "Apache Hadoop." http://hadoop.apache.org/.

The Cancer Genome Atlas. 2008. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways." *Nature* 455 (7216) (October 23): 1061–8. doi:10.1038/nature07385.

The Cancer Genome Atlas Network. 2012. "Comprehensive Molecular Characterization of Human Colon and Rectal Cancer." *Nature* 487 (7407) (July 19): 330–7. doi:10.1038/nature11252.

The International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426 (6968) (December 18): 789–796.

Thompson, Sarah L, and Duane A Compton. 2010. "Proliferation of Aneuploid Human Cells Is Limited by a p53-Dependent Mechanism." *The Journal of Cell Biology* 188 (3) (February 8): 369–81. doi:10.1083/jcb.200905057.

Treangen, Todd J., and Steven L. Salzberg. 2012. "Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions." *Nature Reviews Genetics* 13 (November 2011). doi:10.1038/nrg3164.

Vanden Berghe, Tom, Andreas Linkermann, Sandrine Jouan-Lanhouet, Henning Walczak, and Peter Vandenabeele. 2014. "Regulated Necrosis: The Expanding Network of Non-Apoptotic Cell Death Pathways." *Nature Reviews. Molecular Cell Biology* 15 (2) (February): 135–47. doi:10.1038/nrm3737.

Visvader, Jane E. 2011. "Cells of Origin in Cancer." *Nature* 469 (7330): 314–322. doi:10.1038/nature09781.

Vogelstein, Bert, and Kenneth W Kinzler. 2004. “Cancer Genes and the Pathways They Control.” *Nature Medicine* 10 (8): 789–799. doi:10.1038/nm1087.

Weber-Mangal, Susanne, Hans-Peter Sinn, Susanne Popp, Rüdiger Klaes, Robert Emig, Martin Bentz, Ulrich Mansmann, Gunther Bastert, Claus R Bartram, and Anna Jauch. 2003. “Breast Cancer in Young Women (< or = 35 Years): Genomic Aberrations Detected by Comparative Genomic Hybridization.” *International Journal of Cancer* 107 (4) (November 20): 583–592. doi:10.1002/ijc.11460.

Wei, Junping, Mark Wunderlich, Catherine Fox, Sara Alvarez, Juan C Cigudosa, Jamie S Wilhelm, Yi Zheng, et al. 2008. “Microenvironment Determines Lineage Fate in a Human Model of MLL-AF9 Leukemia.” *Cancer Cell* 13 (6) (June): 483–95. doi:10.1016/j.ccr.2008.04.020.

Welch, John S, Timothy J Ley, Daniel C Link, Christopher A Miller, David E Larson, Daniel C Koboldt, Lukas D Wartman, et al. 2012. “The Origin and Evolution of Mutations in Acute Myeloid Leukemia.” *Cell* 150 (2) (July 20): 264–78. doi:10.1016/j.cell.2012.06.023.

Wong, Kim, Thomas M Keane, James Stalker, and David J Adams. 2010. “Enhanced Structural Variant and Breakpoint Detection Using SVMerge by Integration of Multiple Detection Methods and Local Assembly.” *Genome Biology* 11 (12) (January): R128. doi:10.1186/gb-2010-11-12-r128.

Ye, Kai, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 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) (November 1): 2865–71. doi:10.1093/bioinformatics/btp394.

Zack, Travis I, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, et al. 2013. “Pan-Cancer Patterns of Somatic Copy Number Alteration.” *Nature Genetics* 45 (10) (September 26): 1134–1140. doi:10.1038/ng.2760.

APPENDIX

PAPERS PUBLISHED

Killcoyne, S. & del Sol, A. **Identification of large-scale genomic variation in cancer genomes using in silico reference models.** *Nucleic Acids Res.* gkv828 (2015). doi:10.1093/nar/gkv828

Killcoyne, S. & del Sol, A. **FIGG: Simulating populations of whole genome sequences for heterogeneous data analyses.** *BMC Bioinformatics* 15, 149 (2014).

Nucleic Acids Research Advance Access published August 11, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv828

Identification of large-scale genomic variation in cancer genomes using *in silico* reference models

Sarah Killcoyne and Antonio del Sol*

Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, 6, Avenue Swing, Belvaux L-4367, Luxembourg

Received January 29, 2015; Revised July 10, 2015; Accepted August 01, 2015

ABSTRACT

Identifying large-scale structural variation in cancer genomes continues to be a challenge to researchers. Current methods rely on genome alignments based on a reference that can be a poor fit to highly variant and complex tumor genomes. To address this challenge we developed a method that uses available breakpoint information to generate models of structural variations. We use these models as references to align previously unmapped and discordant reads from a genome. By using these models to align unmapped reads, we show that our method can help to identify large-scale variations that have been previously missed.

INTRODUCTION

Cancer genomes are diverse, and often differ considerably from their germlines. This means that the standard *reference-based* mechanisms used in genome alignment are not always suitable. These mechanisms rely on the assumption that the reference is highly similar to the sample genome. As the reference being used is not an accurate representation of the cancer genome(s), alternative strategies that can use references that represent large-scale structural variation are needed. This paper introduces one such method that uses prior information about known characteristics of cancer genomes to inform a search strategy, which allows for a more efficient mapping of reads against alternative references.

One of the problems with cancer genomes is that they exhibit a high degree of structural variation from the germline. Genomic structural variation is defined as alterations to the genome sequence such as duplication, copy number variation, inversion or translocation (1). While the size of small structural variants can range from anything over a single base pair to 1kb, large-scale variations can involve up to several million base pairs and result in chromosomal aberrations that can be seen at the microscopic level.

Prior to the advent of high-throughput sequencing (HTS) technologies, microscopic methods enabled the identification of cancer structural variation at the level of chromosomal aberrations. Large insertions, deletions or translocations could be identified in a karyotype using Giemsa staining, fluorescence-*in situ* hybridization (FISH), or spectral karyotyping (SKY), and associated with disease phenotypes. These large-scale chromosomal aberrations are rare in the population generally (due to developmental lethality in most cases) and are often associated with severe disease phenotypes. However, the number and complexity of these large variants can be high in tumor genomes (2–4).

In a number of cancers these microscopic levels of structural variation are clinically significant markers of tumor type and malignancy. Known variants can be used to stratify a patient's disease as in multiple myeloma with recurrent translocations between chromosomes 4, 11 and 14 (5), while others such as the Philadelphia chromosome in leukemia (chronic myelogenous and acute lymphoblastic) results in a clinically significant gene fusion BCR/ABL1 (6), which is used in targeted drug therapy (7). Additionally, it has been shown that mutational complexity, including chromosomal aberrations, increases over time contributing to an aberrant activation/repression of multiple genes and therefore potentially contributing to drug resistance or metastasis (8–10). This means that while many translocations (both intra- and inter-chromosomal) have been identified, an individual patient's tumor genome could display a complex mixture of structural variations which may not already be characterized.

As sequencing has become a common method of identifying individual variants in both clinical and research labs, identifying large-scale variants from HTS data alone is increasingly important. There are still a number of issues in identification of large-scale variants in short-read sequence data. The first issue is due to the small size of reads relative to the variation. The current generation of HTS technologies were developed to enable the rapid sequencing of entire genomes through the parallel sequencing of overlapping short-reads (11). In pair-ended HTS short segments are sequenced (e.g. 35–250 bp for Illumina) from two ends of a fragment of known length (e.g. 200–800 bp for Illumina). When aligning these reads to a reference the insert size between each read pair allows the alignment algorithm to in-

*To whom correspondence should be addressed. Tel: +352 46 66 44 6982; Fax: +352 46 66 44 6949; Email: antonio.delsol@uni.lu

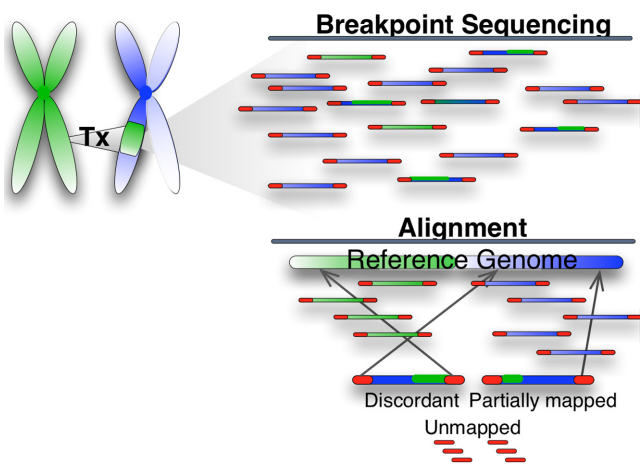


Figure 1. Intra- or Inter-chromosomal translocations result in read pairs that will not align with the expected insert size between the pair. Some of these may be aligned, but will lack information regarding the mapping quality, as that is dependent on the gap and location information. Many reads resulting from this sort of variant may also be unaligned. These issues are a direct result of the use of a reference sequence that does not reflect the structure of the sample sequence.

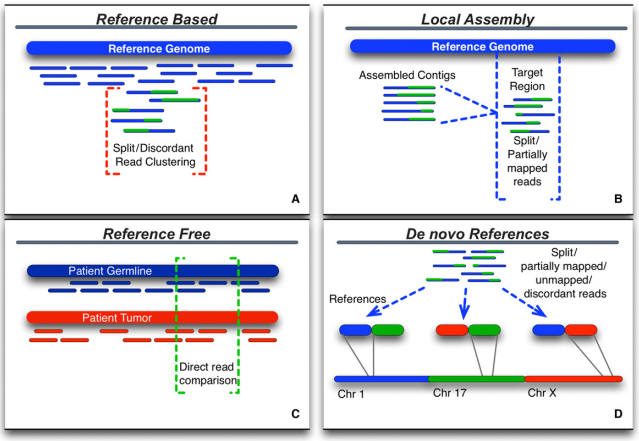


Figure 2. Structural variation detection methods have generally used the *reference based* (**A**) approach where reads are first aligned to the reference, then clustered using a variety of measures to identify SVs. Most of these methods suffer from the assumption that the alignment of the read that is reported is the correct one, due to computational limitations of the alignment algorithms. Alternative methods proposed include performing a local realignment (**B**) of misaligned reads after building longer contigs of the putative breakpoint regions (e.g. *Local Assembly*), and directly comparing the reads of two or more genomes (**C**), in this case tumor/normal, to identify mutations ranging from single nucleotide to inter-chromosomal translocations (e.g. *Reference Free*). Our method (**D**) aims to identify breakpoint regions by generating multiple small references which model potential breakpoints (e.g. *De novo References*), complementing the existing methods.

dicate that a read-pair is correctly aligned to a specific location (12). However, these locations are dependent on finding a good alignment to the provided reference, and in the case of a cancer genome with large-scale structural variation such a reference will be a poor fit (see Figure 1).

In large-scale variations a breakpoint and recombination occurs at a potentially fragile location on the chromosome, altering the sequence. A read-pair generated from this genome can span the breakpoint (if it happened to fall within the gap of the read-pair), or result in a 'split' read where the beginning and end of the read align to different locations. However, due to computational limitations inherent in sequence alignment many reads that could identify these breaks may not be mapped to the reference (13,14).

Complicating the already difficult task of identifying large-scale structural variants in tumor samples is the high degree of genomic heterogeneity present. Samples taken from a solid tumor can include multiple sub-clonal cellular populations that do not share the same variants (15). The result in a sequencing sample is a low frequency of reads supporting a given variation, and in large structural variants some or all may also be unmapped and therefore unavailable for identification.

These difficulties have resulted in a variety of methods being developed which use short-read sequencing data to identify structural variants. The most commonly used methods are *reference based* (see Figure 2A) where variant analysis relies on the initial alignment of sequence reads to the reference genome. When using the aligned reads the existence and position of a breakpoint, and the resulting structural variation, is inferred through clustering or windowing strategies (16). The 'discordant' reads (e.g. mapping to different chromosomes or with incorrect orientation) are used to identify a possible breakpoint through clustering the reads as in BreakDancer (17) or Pindel (18). While PRISM (19), DELLY (20) and SoftSearch (21) cluster 'split-reads' where one of the pair has mapped unambiguously to the reference genome or the CIGAR value has significant numbers of soft-clipped bases (e.g. partially aligned reads). As these methods often limit the size of the variants they can detect consensus approaches such as SVMerge (22) are often used to increase detection across all types.

While the *reference based* approach is widely used, it suffers from a number of limitations inherent to current alignment algorithms, and these methods are therefore unable to identify many structural variants in highly heterogeneous samples. These limitations include the current short length of read-pairs along with the highly repetitive sequence of the human genome. This makes it highly likely for a read pair to align to multiple locations (23,24) across the genome. Alignment algorithms most often use the 'best mapping' approach to reporting read alignment, where the alignment resulting in the fewest mismatches is reported or when all are equally good matches one is randomly selected (25). This is due both to algorithmic constraints, as allowing mismatches increases the number of possible alignments, and to simplify downstream computation as the reads with the highest quality scores are used in variant detection (26). *Reference based* alignment algorithms (e.g. BWA (27), Bowtie (28), SOAP2 (29)) cannot practically do an exhaustive search in the case of reads which may have multiple alignments (as can happen with high rates of variation or with too many mismatches), and where alignments are found will typically only report one of many possible alignments. This is a fundamental limitation to methods that rely on reference alignment, especially when considering tumor genomes. Thus alternate methods that rely less on the reference genome are now being developed.

In the last two years, two alternative approaches have been published which do not rely directly on the standard reference genome alignment. The first approach is best described as a *Local Assembly* (see Figure 2B) method, developed by Abo et.al. (30). This method reassembles misaligned reads within a target specific region into contigs using an overlapping kmer seed from the sequence reads and target region. These contigs are then realigned within the target regions and classified into specific variant types (e.g. inversions, indels, translocations). Locally assembling contigs from regions that have high rates of misalignment overcomes one of the major issues inherent to short-read sequencing, namely that the read lengths are too short to uniquely align when the genome has been structurally altered. While this approach cannot currently scale genome-wide, as it effectively involves *de novo* assembly, it is ideal for resequencing experiments or targeted identification in whole genome or exome data.

The second alternative is the *reference free* (see Figure 2C) method, which takes a completely different approach and avoids the reference entirely by directly analyzing the reads without first aligning them. In this case there is no positional information known, and here the methods vary widely in their implementation. Hormozdiari et.al. (31) assumes that structural variants can be detected with higher accuracy by using multiple related genomes. In this case while a reference genome is used as an intermediary in the analysis, the authors assume that the true variants are discoverable by simultaneously comparing patient genomes directly. They show this clearly with small structural variants (<1 kb) in several genomes from the YRI population in the 1000 Genomes data (32) and a family trio, though it is less clear how well this may work in complex tumor samples. A more recent *reference free* approach called SMUFIN (33) directly compares reads without alignment and was developed specifically for the tumor/normal pairs of genomes. Here it is expected that reads will be highly similar and mutations can be found by grouping reads into a tree structure that branches where mutations are found. Breakpoints can be identified in the branches of the read tree, and local alignment performed. Both *reference free* approaches identify structural variation with greater accuracy than the primary *reference based* approaches.

The analysis for SMUFIN also showed that there might be significantly more complex large-scale structural variation in tumor sequences than has been previously reported. This is due in large part to the fact that tumor genomes can include highly complex low frequency variations and the reference genome that alignment algorithms rely on cannot model these in mixed samples. The methods that rely on the *reference based* alignment (e.g. BreakDancer, SoftSearch, etc.) are limited by the aligners and, as is shown by both the *local assembly* and *reference free* methods, alternative approaches are necessary to overcome the alignment issues.

Here we propose a third alternative for identifying structurally variant regions, which can complement the existing methods in complex tumor samples: *de novo* generation of multiple references. Our *de novo* method (see Figure 2D) generates a large number of new references that model potential structural variations. We use a tuned optimization strategy based on prior information from karyotypes across many different cancers to select suitable references. Standard alignment tools are then used to align previously unmapped and discordant read-pairs to the new references, and the resulting alignments are scored. Using high-performance cluster computing this process can be repeated hundreds of times to select likely breakpoint recombination regions.

In the *Material and Methods* section, we describe our strategy starting with the generation of *de novo* references followed by identification of regions that may include structural variations. In the following *Results* section, we show that our identification method can find structural variants in simulated data and then we apply it in several patients from The Cancer Genome Atlas (TCGA) (34). Finally the *Discussion* section discusses the need for alternative strategies to identify structural variations in tumor samples and specifically the advantages and limitations of our *de novo* method.

MATERIAL AND METHODS

The strategy we have used to enable more accurate alignment of cancer genomes from short reads is to limit the search space in which reads can align by altering the reference (see Figure 3), and therefore decreasing the read distance between potential split read-pairs.

In order to limit the search space while using standard alignment tools, the reference is replaced by a series of *in silico* reference sequences that model chromosomal recombination regions. The generation of *in silico* references requires a pre-populated database of breakpoint frequencies, including chromosomal locations generated from available breakpoint data, and are used to align only those reads in a sample that were previously partially or fully unmapped or discordant.

In silico model generation

Instead of aligning against a single reference, our method aligns against hundreds of smaller references. These smaller references model potential structural variations seen in cancer originating from fragile regions in the genome. The set of new references contain sequences from two different genomic regions thus simulating the result of a recombination event. These models are generated using prior knowledge of breakpoint frequencies in cancer based on karyotype data (e.g. breakpoints at cytogenetic bands). These frequencies were obtained from analysis of public karyotype data sets including patient karyotypes and cell lines:

- ***Patient karyotypes***. 99 764 across many different (poorly curated) cancer types were analyzed from the Mitelman CGAP database (35) and 325 from NCI and NCBI's SKY/M-FISH and CGH Database (http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi). The majority of these were blood cancers (e.g. leukemia, lymphoma and myeloma).
- ***Cell line karyotypes***. 84 were analyzed from the University of Cambridge CGP SKY/FISH of Epithelial Cell Lines (http://www.path.cam.ac.uk/~pawefish/) and 67 from the NCI Fredrick National Laboratory NCI60

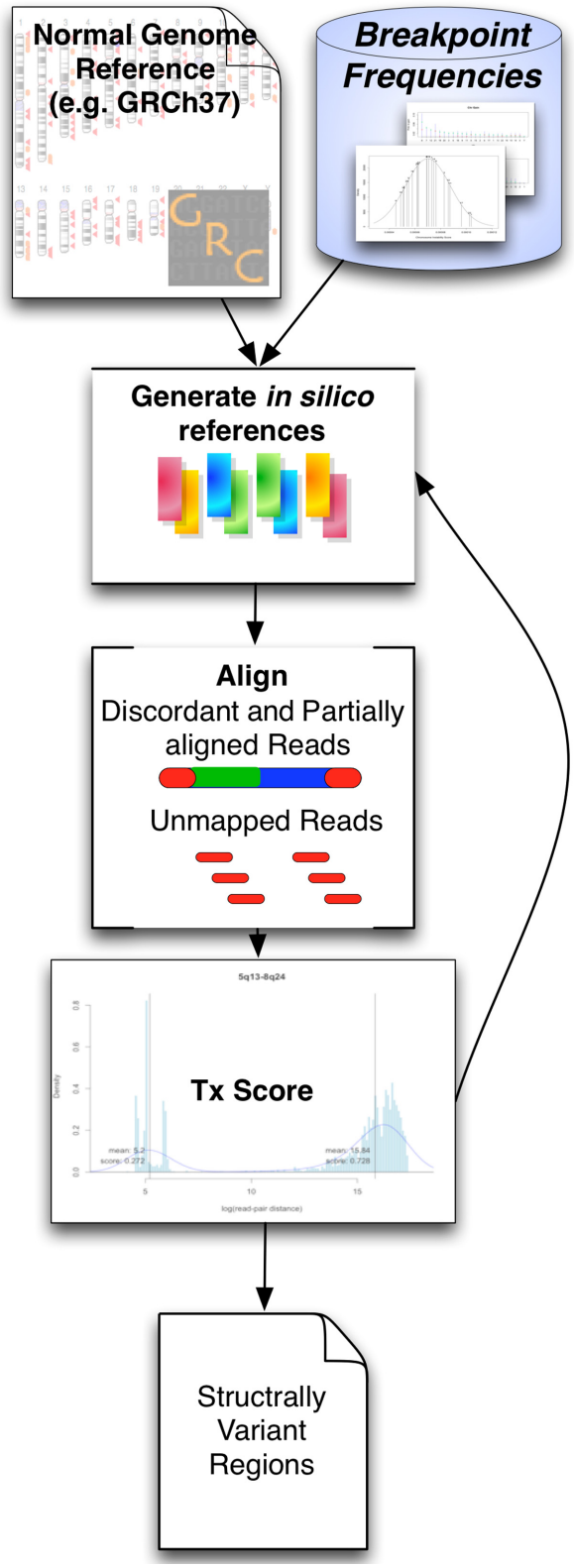


Figure 3. The general workflow of this method takes prior knowledge of the likelihood of an aberration at points in the chromosome to generate a new *in silico* reference. This is then used to align the discordant and unmapped reads from a previously aligned sample, scored and clustered to identify the best scoring regions.

Cell Line Drug Discovery Panel (http://home.ncifcrf.gov/CCR/60SKY/new/demo1.asp). These were curated simply based on the tissue involved (e.g. 'heart' or 'thymus').

There are 320 major cytogenetic bands within the human genome (36), all of these are found to be involved in at least one breakpoint reported within the available data sets. A pairwise combination of each of the bands to create simulated references results in $C_{320,2}$(51 040) possible combinations. While this would not seem to be too many combinations to test, and ideally testing against all of them would offer the most comprehensive view, there are computational limitations. First is disk space: the index for all simulated reference combinations requires 2.5 TB of hard disk space, and the subsequent alignment BAM files for a small set of reads (1.9 million) from a single genome would require more than 30 TB on disk. The second limitation is the alignment step itself. Aligning a small number of reads against many smaller references is an ideal situation for parallelization, however each single alignment (e.g. bwa mem –a –t 12) plus analysis computation still required 65 min on a single node in our local cluster. All 51 040 pairwise regions would require 840 000 compute hours (or 96 years) in order to align and analyze. Therefore even with access to a HPC cluster and a high degree of parallelization, this is a computationally intensive method.

Instead of using all possible *in silico* references we use an informed search strategy. This informed approach is required to select regions that should be tested for breakpoints. As each breakpoint is not equally likely based on the karyotypes described, and to further decrease our search space and computational load, the frequencies calculated from karyotype data are used to generate a set of several hundred simulated references. This informed approach is outlined in the *Optimisation of Reference Selection* section below.

Structural variant detection

The generated references now act as model regions for possible large-scale structural variation. Limiting the search space by creating smaller references also allows us to increase the number of possible alignments by including previously unmapped reads. A filtered set of reads from a patient sample that includes only those reads that were already aligned to different chromosomes ('discordant') or where one or both reads were unmapped are then aligned to these smaller references in parallel. This enables the method to rapidly compare multiple possible recombination regions. As we have limited the search space by using a smaller reference we are able to relax the search criteria to allow for more exhaustive searching and greater mismatches.

In each model region the aligned reads are filtered to limit the inclusion of poor quality data. As these reads were previously unmapped, we filter out reads from the alignments that are below the mean summed Phred quality score identified from the original BAM. Additionally, any alignments where 50% or less of the read have matched according to the CIGAR (see SAM format) value are discarded. Each model region is then evaluated by analyzing the distribution of read-pair insert sizes in each new

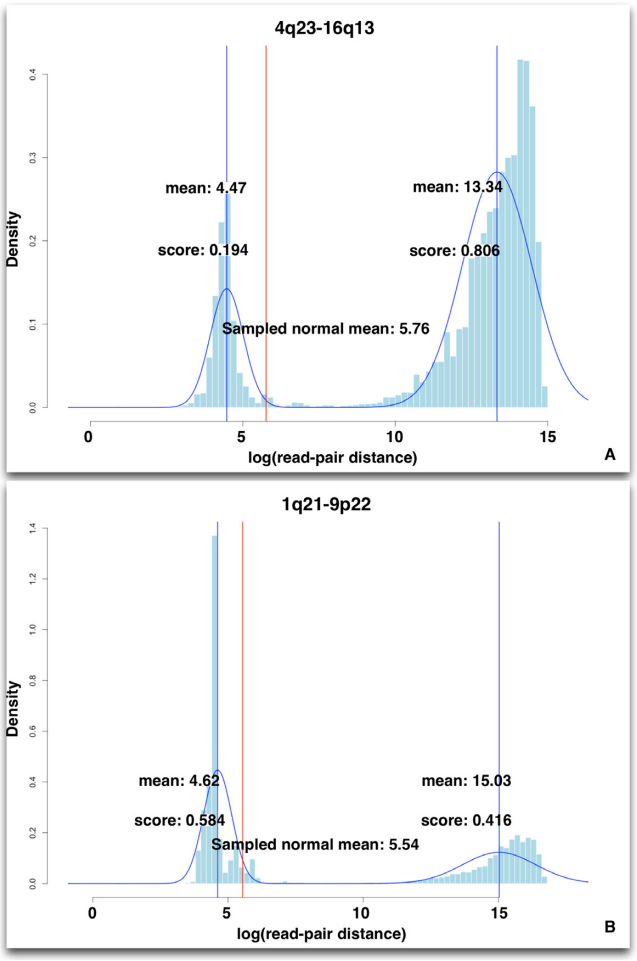


Figure 4. These plots show representative distributions of logged read-pair insert sizes from aligned reads against two different *in silico* references. Both aligned references show a bimodal distribution with two clear centers, however (**A**) shows an alignment to a reference that may be representative of a SV in the sample, while (**B**) is representative of an alignment that shows only noise based on the EMr values for the second distribution in each plot. In both (**A**) and (**B**) the first distributions (on the left of each plot) are from read-pairs that align with an insert size <2 s.d. from the mean, while the second distributions (on the right of each plot) are from read-pairs that align with an insert size >4 s.d. from the mean. In (**A**) the aligned reads in the second distribution show a very clear signal with a large number of supporting reads, suggesting there is a SV within this region. In (**B**) the second distribution is poorly defined. The reads in this region are less likely to indicate a SV as the second distribution does not exceed the noise from the first.

alignment. A bimodal distribution of the logged insert size between aligned read-pairs is observed across the *in silico* reference alignments. In all model regions, the distributions are consistently bimodal and non-symmetric, and each peak is reflective of these two different possible alignments (see Figure 4). We find these sub-distributions using the Expectation-Maximization (EM) algorithm (R package 'mclust' (37)). The first distribution is characterized by reads with a small insert size (<2 s.d. of the mean insert size) and which are poorly mapped, having a map quality score <30. This indicated that the first distribution is basically noise, and arises due to alignments with a high likelihood of error. The second distribution includes reads that consistently align with an insert size that was >4 s.d. from the mean insert size, as used in (16). Due to insert size there is no map quality score, but we can overcome this by including only those alignments with higher CIGAR and Phred values.

The individual *in silico* reference score is calculated in two parts. The first is based on the mixture model parameters for the second distribution as calculated by EM:

$$EMr = \frac{\sum_{n=1}^{N} P(n|z)}{N}$$

Where $P(n|z)$ is the conditional probability of the nth read belonging to each of the two distributions identified. The EMr reflects the proportion of reads that are found to have a higher mate pair distance, and is derived by finding the probability of the nth read belonging to the second distribution, then iterating over the set of N where N is the total number of reads aligned to this reference. The resulting value is a ratio based on the number of distributions found and the sum of the EMr for each is 1. As the first distribution describes 'noise' in the alignments we can use it to find a cutoff value for further analysis of the alignments described by the given *in silico* reference. All models where the second distribution have an EMr below the cutoff can therefore be discarded.

The second part of the score is based on a sliding-window clustering approach to identify breakpoint locations based on alignment positions. Discordantly aligned reads from the second distribution are clustered by position if the read pairs also span both chromosomes represented by the simulated reference. This provides an estimation of windowed depth-of-coverage as discussed in (16) for a translocation breakpoint. However, this is not meant to provide a direct analysis of the breakpoint location, instead this provides a necessary adjustment for the EM score in the second distribution above.

$$Tx = EMr + \frac{W_{max}}{N_b}$$

Where W_{max} is the cluster with the highest total count of reads from the second distribution, and N_b is the total number of reads within the second distribution.

Simulated data sets

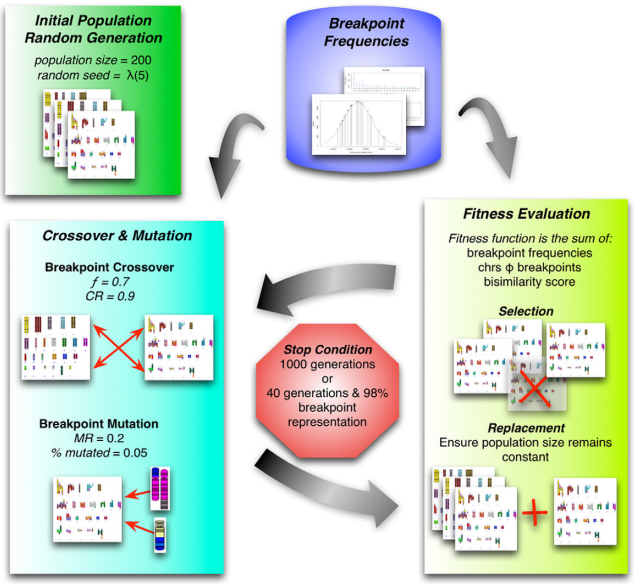
In order to estimate sensitivity and specificity for the Tx scores and subsequent structural variant calling, we generated reads using the ART (38) read simulator for Illumina in 20 sets of randomly selected pairs of chromosomes and cytogenetic bands. Each set included a randomly selected inter-chromosomal translocation based on position and sequence information from genome assembly GRCh37 (see Supplementary Table S1). The only limitation placed on the simulated breakpoints was that they did not fall directly on a cytogenetic band boundary and that they were in a region that could be aligned (e.g. avoiding poorly sequenced or highly repetitive regions such as most centromeres and telomeres). The analysis of these data is discussed in the *Results* section.

Optimization of reference selection

As noted above, there are tens of thousands of pairwise combinations possible for just the major cytogenetic bands. Using probabilities to generate the most likely combinations will result in identifying reads that belong to well known breakpoints, while missing those that are less well characterized or are unreported in the literature. This means that in order to optimize the selection of simulated references, and avoid bias toward the most commonly known breakpoints (e.g. centromeres are the most reported breakpoints in the microscopic methods, or the Philadelphia chromosome in leukemia), a selection algorithm is introduced to generate populations of breakpoints. These populations are generated as individuals with full chromosomal complements and aberrant chromosomes.

This selection uses a type of genetic algorithm known as differential evolution (DE) (39) with an optimization function for the entire population being iterated over, instead of a single solution. This function combines the fitness of all individual references, and a measure of the diversity (see Supplementary Methods) of the DE population (see Figure 5). The diversity score ensures that cytogenetic bands with a smaller probability of being involved in a recombination event may be represented, enables the testing of chromosomal regions that may otherwise be underrepresented due to a bias in the frequency data (e.g. missing data for disease-specific aberrations), and avoids over-testing breakpoints that may be overrepresented in the knowledgebase (e.g. centromeres, Philadelphia chromosome, etc.).

The output of the selection algorithm is a population of pairs of chromosomal locations to be used in generating FASTA files. Each of these represents the sequence of the selected recombination. For example t(16;8)(q13;q24), is defined as starting with 16q13 (56700001- 57400000) and ending with 8q24 (117700001- 46364022) creating a recombination point at 700 kb.




```
-- Initialize Population(P) --
P = 200 random Candidates /* Poisson distribution with random seed λ = 5 */

while ( generations < 1000 OR /* Termination conditions */
      (generations > 40 AND breakpoint representation < 98%) ) {
   -- Perform Crossover F(0.7), CR(0.9) --
   foreach (Candidate)  {
      foreach (Breakpoint) {
         /* 3 candidates are randomly selected for each Breakpoint */
         if (random num <= CR) {
            x1,x2,x3 = candidate[1,2,3] has breakpoint? 1: 0
            /* add the differences with weighting factor F */
             v1 = x3 + F * (x1 - x2)
              if (v1 >= 0) add Breakpoint to Candidate
              else remove Breakpoint from Candidate
          }
       }
   }
   -- Perform Mutation  MR(0.3), PP(0.05) --
   foreach (Candidate) {  /* Randomly select P*PP Candidates to mutate */
        if (random num <= MR) add random 1..n breakpoints to Candidate
   }
   -- Select Candidates --
   foreach (Candidate in P) {
      /* Due to the weighting factor of the chr/bp ratio,
         a low value is higher fitness */
       if (Candidate fitness < max fitness) select Candidate

       /* NCD determines pairwise similarity,
          used to ensure population diversity */
       if (CandiateNCD >= MaxNCD) select random Candidate from any similar pair
   }
   -- Add new Candidates to Population --
   while (P < 200) {
       add random new Candidate to P
   }
   generations += 1
}
```

Figure 5. The selection algorithm is an implementation of differential evolution as this variant of genetic algorithms provide multiple solutions across the search space. The process of DE can be summarized as: (**i**) generate initial population (**ii**) cross each breakpoint pair by exchanging partners given a crossover constant (*CR*) (**iii**) mutate each breakpoint pair given a mutation constant (*F*) (**iv**) evaluate the individual fitness (**v**) evaluate the population diversity. When either the population diversity reaches a reasonable optimum or a certain number of generations have been run the selection algorithm stops. The parameters *CR*, *F* and maximum generations were all selected to optimize the diversity of the end population. Each of these constants can also have a large impact on the computational time it takes to generate a population.

RESULTS

We performed a test on simulated data to validate the method and identify suitable parameters for cancer variant selection. Then we applied the parameters we learned from the simulated validation tests to tumor/germline data sets from TCGA directly. For the patient data sets we compare our method to BreakDancer, as it continues to be the most commonly used tool for large-scale variant detection in tumors.

Simulated inter-chromosomal breakpoints

To validate that our method could identify inter-chromosomal breakpoints with a reasonable degree of accuracy we used the simulated data described in the *Methods* section. In the fully simulated set of 20 chromosome pairs we wanted to validate that the inserted breakpoint does result in a high *Tx* score when compared to a set of randomly selected references representing other potential breakpoint alignments. We applied k-means clustering to identify the set of regions with high *Tx* values. In order to keep the false positive rate (FPR) consistently below 10% in subsequent analyses we perform clustering using 4 centroids (see Figure 6).

As our method does not try to select a single unique alignment for each read, we expect to find a higher number of possible structural variations and therefore select a stricter

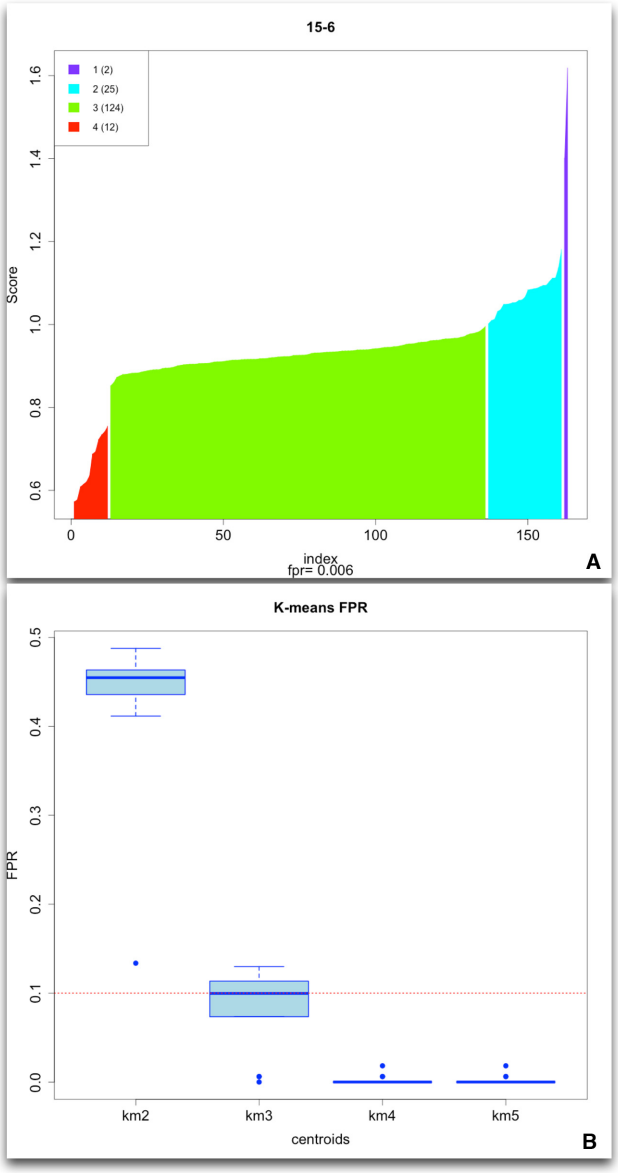


Figure 6. The *Tx* scores reported by our method (**A**) showed a one-tailed distribution. Using k-means we can identify the set of regions with the high *Tx* scores as shown in the purple cluster. This cluster consistently included the simulated breakpoint in all data sets across multiple coverage tests. From the 20 simulated data sets we determined that a selection based on 4 clusters (**B**) (seeding the centroids with quartiles), resulted in a FPR between 5% and 10%.

cutoff to avoid increasing the likely errors. Additionally, the weighting value of the clusters in the *Tx* score $\left(\text{e.g } \frac{W_{max}}{N_b}\right)$ becomes more important in lower coverage or poorer quality simulations. This provides a useful parameter for investigation of structural variants, which are found in smaller sub-populations within the tumor.

In each of the 20 simulated data sets clustering the *Tx* scores resulted in identification of the known breakpoint in the top cluster of each set. It also enabled calculation of FPR values based on the clusters. These values were used in subsequent analysis of the TCGA patient data.

Analysis: detection of large-scale variants in TCGA patients

We applied our method to the analysis of nine different matched tumor/germline genomes from TCGA across seven different cancer types (see Supplementary Table S2). Each set of unmapped and discordant reads from the genomes was compared against the same set of 278 regions selected by the optimization algorithm (see Supplementary Table S3). Based on the FPR rate calculated in the simulated data set above, we used a result selection from k-means clustering and took only the top cluster for analysis. We then filtered the germline hits from the tumor list in order to compare with BreakDancer. In each patient we identified regions that score highly for breakpoint inclusion. By using the germline samples to filter the results of the tumor samples we were able to remove regions that appeared to have significant unspecific read alignments as they were often found in both tumor and normal tissue samples.

Two patients had been previously analyzed as part of a large cohort study for TCGA: the glioblastoma (GBM) patient (40) and one of the colon/rectal (COAD (2)) patients (41). In these two patients we found no regions that were highly scored over their germline, which was consistent with earlier analyses. Our analysis of the GBM patient found no somatic structural variation, which was consistent with the cohort analysis where no structural variants were found in this patient. Our analysis of the COAD patient also found no structural variations. This was also consistent with the original COAD analysis as this patient was not found to have any structural or copy number variations.

It is also worth noting that in the ovarian patient (OV) samples our method identified two bands that are found as part of multiple regions (9q13, 4q13), which may suggest complex rearrangements. The 9q13 band is a known fragile site that is commonly involved in pericentric inversions in the germline linked with ovarian cancer (42), while 4q13 has been found to have a high rate of copy number variation in BRCA1 associated ovarian cancers (43).

***Reference based* comparison**

We compared these results to the BreakDancer analysis of tumor/germline pairs (see Table 1). Each translocation identified by BreakDancer was mapped back to the corresponding chromosomal region. Several findings are important here:

- ***Commonly identified regions.*** Structural variations that are found commonly across cancer types are highly likely to be due to biases in the detection method rather than a set of rearrangements that are common across cancer types. Both BreakDancer and our *de novo* method are going to find these due to the reporting of aligned reads. However, within the highest scoring regions, BreakDancer tends to find breakpoints in the same regions across multiple patients and cancers. Across all 32 regions, 26 were identified in more than one patient. For instance in the top regions with the highest scoring breakpoints, translocations in the 1p11–17p11 region were found in 6 of the 9 patients, while breakpoints in 1p11–11p11 are found in 4. In fact in the COAD-2 patient, where our method found no difference between

Table 1. Shows the regions in which BreakDancer and our *de novo reference* method detect large scale structural variations

BRCA (1)	BRCA (2)	COAD (1)	COAD (2)	GBM	KIRC	LUAD	OV	LAML-14
Reference based: BreakDancer								
1p11–19q11	1q12-Yq11	1q12-Yq11	1q12–21p11	**1p11–17p11**	**1p11–17p11**	**1p11–19q11**	1p22–17q12	**1p11–17p11**
	1q21-Yq11	1q43–10p11	1p34–6p22		1q21–4p11	**1p11–6p11**	1p22–9q22	**1p11–11p11**
	1p11–17p11	1q21-Yq11	1q21–16p11		1q12–21p11		**1p11–11p11**	
	1q21–4p11	1q21–4p11	**1p11–17p11**		**1p11–11p11**		1q21–16p11	
	1q21–16p11	1q21–21p11	3q27–6q15				**1p11–17p11**	
			1p11–11p11				1p34–6p22	
De novo references								
Xq21–9q13	17q23–4q13	3q12–8p11			3q12–8p11	11p15–12q11	3q12–8p11	4q23–16q13
	4q13–2q14	11p15–12q11				6q22–11q21	10p14–9q13	3q12–8p11
		10p14–9q13				Xq21–9q13	17q23–4q13	2q13–14p13
		7p11–13q14				6q15–10p13	22q13–9q13	3q27–6q15
		Xq21–9q13				21q22–14q22	5q32–9q13	17p11–13p12
						10p14–9q13	4q13–2q14	5q32–9q13
						9p22–14q21		
						5q33–9p23		

The table shows representative results from nine different patient samples, all from TCGA. Regions in bold are those that occur in more than one patient, and regions in red under the breast cancer sample are those that are potentially erroneous (as they include alignments in the Y chromosome, where the clinical information list this patient as female). BreakDancer results show an overrepresentation of structural variations in centromeric regions in their top scoring translocations. The bolded regions all include the centromere 1p11, which is poorly sequenced with 80% of the bases missing in the current assembly. The *de novo references* method also results in a few shared regions and centromeric regions (e.g. 3q12–9p11 shared in 2 patients), however, it also finds more regions that include structural variation in the gene rich regions of the genome.

tumor/germline, BreakDancer found only one region (3q27–6q15) where both bands involved were not identified in any other patient. Our method will also find regions that are common across cancers if the same set of regions are tested, however it is less likely with only 7 of the 29 regions being found in more than 1 patient, and the most common one (3q12–8p11) found in 4.

- ***Centromeres overrepresented.*** Across all 9 patients and cancers, centromeric regions are overrepresented in the top scoring breakpoints found by BreakDancer with 27 of the 32 regions including at least one centromere. Furthermore, in 7 of the patients all of the top 20 identified regions include at least one centromere. As centromeres (e.g. q11 and p11) make up only 15% of the major cytogenetic bands we would expect to see only 10 regions in 32 include a centromere in an unbiased sampling. As with the commonly identified regions, this is unlikely to be due to a common cancer event. Centromeres are poorly characterized across the chromosomes due to their highly repetitive sequences (25). This makes it more likely that reads aligning within a centromere will have correct alignments elsewhere in the genome. In fact the region shared across most patients (e.g. 1p11, found in 40% of the top regions) is poorly sequenced with 80% of the bases lacking a known assembly. Comparatively, our *de novo* method finds only 8 centromeres in our top 29 regions.
- ***Inaccurate alignments.*** Of the regions identified by BreakDancer with the highest scoring translocations in BRCA (breast cancer) two include alignments to the Y chromosome, however associated clinical data lists this patient as female. While this is not impossible, it would suggest issues with the alignment. As our method is not relying on a single reported alignment we did not find a similar inaccuracy.

The highly duplicative nature of the regions that are the most commonly found in translocations by BreakDancer suggests that only very common breakpoints are found, and that there are issues with alignment. Alignment algorithms typically report only a single 'best' alignment leaving BreakDancer, and other *reference based* methods, with limited information on which to make identifications in complex samples. This is not to claim that the identifications are necessarily incorrect. Centromeres are likely to be involved in various types of large-scale structural variations due to microtubule defects (44), and it is not unlikely that certain regions are 'hot spots' for breakage and recombination. However, as reported by both SMufin and BreaKmer, *reference based* methods (and BreakDancer specifically) are missing large numbers of structural variants due to their reliance on the alignment algorithm.

As our method uses multiple *de novo* references to model potential breakpoints prior to alignment, we find regions with variation that *reference based* methods such as BreakDancer cannot. Thus while *reference based* methods provide a good initial estimate for variations, concurrent use of *reference free* or *de novo reference based* methods can provide a more complete view of the variation present in the tumor.

One of the difficulties present in the analysis of structural variation in cancer, is that in the absence of a directly gene related product (e.g. gene fusion) the effects may be subtle. For instance, if the break and recombination of a translocated segment occurs at intergenic regions the translocation may only affect the regulation or transcriptional enhancement of a gene rather than entirely inactivating the gene or increasing gene expression. Alternatively, as tumor genomes are often made of up highly heterogeneous cellular populations it is also likely that the effect of a fusion occurring in a sub-clonal population is at levels below our ability to detect in gene expression studies. It is therefore worth noting the regions identified by our *de novo* method may not result in fusion genes such as BCR/ABL but in altered regulation, or no detectable change.

In the second breast cancer patient (BRCA (2)) band 4q13 was represented in both of the top scoring regions, suggesting that alignments in that band were the main driver for the high scores. We generated an additional 10 regions that included 4q13 and added them to the pool of regions then performed the clustering again. All regions that included 4q13 were highly scored in the tumor sample, but not the germline. This is important to note for a few reasons. First, 4q13 is one of the regions known to integrate viral DNA

from human papillomavirus (45) suggesting that there may be fragile sites for other types of structural rearrangement (and making it an important region for cervical and ovarian cancers as well). Secondly, several genes important to the development or aggressiveness of breast cancer including EREG, which is involved in ER/HER2 status, are located within this band. Finally, while the top match, 17q23–4q13, has not been reported as a structural variation previously both bands have been identified as showing significant copy number gains in aggressive breast cancers (46).

DISCUSSION

The method introduced in this paper has a number of advantages and limitations. The main advantage is that our method is able to find structural variations that most commonly used tools would be unable to find. The reason these methods are unable to find these structural variations is that they lack the information required to identify them, as an exhaustive search during alignment is computationally inhibitive.

The limitations of our method are due to the fact that the increase in the number of references increases both the noise and search space. Aligning to multiple references results in multiple alignments reported for a single read, increasing the potential for noise. While this is an issue with the method, it is also a necessary condition for the identification of structural variations in complex samples from a tumor. By applying EM and a clustering parameter the regions with the highest likelihood of variation can be selected with a reasonable FPR. Secondly, this is a computationally heavy approach in that there are more possible combinations than can be reasonably tested. However, using prior knowledge and the search optimization algorithm we can limit the search space for each genome tested. This means currently we may miss regions that have structural variation as we do not search all possible regions instead using HPC tools and optimizations to decrease the overall time and computational load. By limiting our search both to the set of unmapped and discordant reads, and using a set of model references based on prior information, we are able to evaluate many possible regions. Ultimately this increases the ability to identify low-frequency aberrations likely to be present in heterogeneous tumor samples.

The identification of large-scale structural variation in cancer genomes continues to be difficult. Most of our current strategies have relied on alignment to a reference that is built on a 'normal' genome, assuming that the sample genome will align well enough for analysis. Unfortunately, due to the potential complexity of large-scale variations, the heterogeneity of a tumor sample itself, and the limitations of short-read sequencing, use of the standard reference can result in poor alignment for large-scale variant regions.

Alternative strategies that rely less on a reference genome, or skip alignment entirely, have provided evidence that the current *reference-based* methods cannot provide a complete view of the range and complexity of structural variation in tumor genomes. Thus, it is important to continue to explore alternative methods for large-scale variant detection in tumor samples. Our method approaches this issue by using multiple references to model potential breakpoints, decreasing the search space in which alignment algorithms function, and overcoming the problem of 'best' alignment mapping by allowing all alignments for each read and evaluating each region individually. This is an important step in tumor analysis due to the presence of clinically significant subclonal populations with complex chromosomal rearrangements.

AVAILABILITY

De novo reference generation is implemented in Java using the Hadoop MapReduce v1.2.1 framework and HBase 0.94. This can be run on a standard desktop machine (without the benefit of parallel computation) with the standalone installation of Hadoop. It may also be run on a cluster which uses Hadoop locally or through Amazon EC2. A compiled version is available at http://sourceforge.net/projects/insilicogenome/files/releases/HBase-Genomes-1.2.jar the corresponding HBase database is available at http://sourceforge.net/projects/insilicogenome/files/Databases/GRCh37.tgz

Analysis of the resulting BAM files is performed in R v3.0.1, available at http://sourceforge.net/projects/insilicogenome/files/releases/denovo_analysis-1.2.tgz

All source code is available on Github (see README files in each module) at https://github.com/skillcoyne/IGCSA

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The computational resources provided in part by the High Performance Computing facility at the University of Luxembourg.

FUNDING

Fonds National de la Recherche (FNR), Luxembourg [4717849 to S.K.]; Amazon Web Services Education Research Grants [PC1MKO8SSJYKK36 to S.K.]. Funding for open access charge: The Doctoral School for Biomedicine, University of Luxembourg.
Conflict of interest statement. None declared.

REFERENCES

1. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
2. Berger,M.F., Lawrence,M.S., Demichelis,F., Drier,Y., Cibulskis,K., Sivachenko,A.Y., Sboner,A., Esgueva,R., Pflueger,D., Sougnez,C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214–220.
3. Quinlan,A.R. and Hall,I.M. (2012) Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.*, **28**, 43–53.
4. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.

5. Garand,R., Avet-Loiseau,H., Accard,F., Moreau,P., Harousseau,J.L. and Bataille,R. (2003) t(11;14) and t(4;14) translocations correlated with mature lymphoplasmacytoid and immature morphology, respectively, in multiple myeloma. *Leukemia*, **17**, 2032–2035.
6. Sattler,M. and Griffin,J.D. (2001) Mechanisms of transformation by the BCR/ABL oncogene. *Int. J. Hematol.*, **73**, 278–291.
7. Schindler,T., Bornmann,W., Pellicena,P., Miller,W.T., Clarkson,B. and Kuriyan,J. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science*, **289**, 1938–1942.
8. Janssen,A. and Medema,R.H. (2012) Genetic instability: tipping the balance. *Oncogene*, **32**, 4459–4470.
9. Baca,S.C., Prandi,D., Lawrence,M.S., Mosquera,J.M., Romanel,A., Drier,Y., Park,K., Kitabayashi,N., Macdonald,T.Y., Ghandi,M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666–677.
10. Stephens,P.J., Greenman,C.D., Fu,B., Yang,F., Bignell,G.R., Mudie,L.J., Pleasance,E.D., Lau,K.W., Beare,D., Stebbings,L.A. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.
11. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
12. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
13. Schbath,S., Martin,V., Zytnicki,M., Fayolle,J., Loux,V. and Gibrat,J.-F. (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J. Comput. Biol.*, **19**, 796–813.
14. Ruffalo,M., LaFramboise,T. and Koyutürk,M. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**, 2790–2796.
15. Greaves,M. and Maley,C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306–313.
16. Medvedev,P., Stanciu,M. and Brudno,M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.
17. Chen,K., Wallis,J.W., Mclellan,M.D., Larson,D.E., Kalicki,J.M., Pohl,C.S., Mcgrath,S.D., Wendl,M.C., Zhang,Q., Locke,D.P. *et al.* (2009) BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
18. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
19. Jiang,Y., Wang,Y. and Brudno,M. (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, **28**, 2576–2583.
20. Rausch,T., Zichner,T., Schlattl,a., Stutz,a. M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
21. Hart,S.N., Sarangi,V., Moore,R., Baheti,S., Bhavsar,J.D., Couch,F.J. and Kocher,J.-P. a (2013) SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One*, **8**, e83356.
22. Wong,K., Keane,T.M., Stalker,J. and Adams,D.J. (2010) Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.*, **11**, R128.
23. Hach,F., Hormozdiari,F., Alkan,C., Hormozdiari,F., Birol,I., Eichler,E.E. and Sahinalp,S.C. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods*, **7**, 576–577.
24. Derrien,T., Estellé,J., Marco Sola,S., Knowles,D.G., Raineri,E., Guigó,R. and Ribeca,P. (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.
25. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
26. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
27. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
28. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
29. Li,R., Yu,C., Li,Y., Lam,T.-W., Yiu,S.-M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
30. Abo,R.P., Ducar,M., Garcia,E.P., Thorner,A.R., Rojas-Rudilla,V., Lin,L., Sholl,L.M., Hahn,W.C., Meyerson,M., Lindeman,N.I. *et al.* (2014) BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res.*, **43**, e19.
31. Hormozdiari,F., Hajirasouliha,I., McPherson,A., Eichler,E.E. and Cenk Sahinalp,S. (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Res.*, **21**, 2203–2212.
32. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
33. Moncunill,V., Gonzalez,S., Beà,S., Andrieux,L.O., Salaverria,I., Royo,C., Martinez,L., Puiggròs,M., Segura-Wang,M., Stütz,A.M. *et al.* (2014) Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.*, doi:10.1038/nbt.3027.
34. The Cancer Genome Atlas. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
35. Mitelman,F., Johansson,B. and Mertens,F. (2015) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. http://cgap.nci.nih.gov/Chromosomes/Mitelman.
36. Kirsch,I.R., Green,E.D., Yonescu,R., Strausberg,R., Carter,N., Bentley,D., Leversha,M.A., Dunham,I., Braden,V.V., Hilgenfeld,E. *et al.* (2000) A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nat. Genet.*, **24**, 339–340.
37. Chris Fraley,A.E.R. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.
38. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
39. Storn,R. and Price,K. (1997) Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.*, **11**, 341–359.
40. Brennan,C.W., Verhaak,R.G.W., McKenna,A., Campos,B., Noushmehr,H., Salama,S.R., Zheng,S., Chakravarty,D., Sanborn,J.Z., Berman,S.H. *et al.* (2013) The somatic genomic landscape of glioblastoma. *Cell*, **155**, 462–477.
41. The Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
42. Yasuhara,T., Okamoto,A., Kitagawa,T., Nikaido,T., Yoshimura,T., Yanaihara,N., Takakura,S., Tanaka,T., Ochiai,K. and Ohtake,Y. (2005) FGF7-like gene is associated with pericentric inversion of chromosome 9, and FGF7 is involved in the development of ovarian cancer. *Int. J. Oncol.*, **26**, 1209–1216.
43. Yoshihara,K., Tajima,A., Adachi,S., Quan,J., Sekine,M., Kase,H., Yahata,T., Inoue,I. and Tanaka,K. (2011) Germline copy number variations in BRCA1-associated ovarian cancer patients. *Genes. Chromosomes Cancer*, **50**, 167–177.
44. Burrack,L.S. and Berman,J. (2012) Flexibility of centromere and kinetochore structures. *Trends Genet.*, **28**, 204–212.
45. Kraus,I., Driesch,C., Vinokurova,S., Hovig,E., Schneider,A., von Knebel Doeberitz,M. and Dürst,M. (2008) The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. *Cancer Res.*, **68**, 2514–2522.
46. Weber-Mangal,S., Sinn,H.-P., Popp,S., Klaes,R., Emig,R., Bentz,M., Mansmann,U., Bastert,G., Bartram,C.R. and Jauch,A. (2003) Breast cancer in young women (< or = 35 years): Genomic aberrations detected by comparative genomic hybridization. *Int. J. Cancer*, **107**, 583–592.

BMC Bioinformatics

SOFTWARE Open Access

FIGG: Simulating populations of whole genome sequences for heterogeneous data analyses

Sarah Killcoyne and Antonio del Sol*

Abstract

Background: High-throughput sequencing has become one of the primary tools for investigation of the molecular basis of disease. The increasing use of sequencing in investigations that aim to understand both individuals and populations is challenging our ability to develop analysis tools that scale with the data. This issue is of particular concern in studies that exhibit a wide degree of heterogeneity or deviation from the standard reference genome. The advent of population scale sequencing studies requires analysis tools that are developed and tested against matching quantities of heterogeneous data.

Results: We developed a large-scale whole genome simulation tool, FIGG, which generates large numbers of whole genomes with known sequence characteristics based on direct sampling of experimentally known or theorized variations. For normal variations we used publicly available data to determine the frequency of different mutation classes across the genome. FIGG then uses this information as a background to generate new sequences from a parent sequence with matching frequencies, but different actual mutations. The background can be normal variations, known disease variations, or a theoretical frequency distribution of variations.

Conclusion: In order to enable the creation of large numbers of genomes, FIGG generates simulated sequences from known genomic variation and iteratively mutates each genome separately. The result is multiple whole genome sequences with unique variations that can primarily be used to provide different reference genomes, model heterogeneous populations, and can offer a standard test environment for new analysis algorithms or bioinformatics tools.

Keywords: Genome sequence, Simulation, Variation frequency, Population

Background

This paper introduces the FIGG (Frequency-based Insilico Genome Generator) tool, which is designed to be of use to computational researchers who require high volumes of artificially generated genomes that mimic the variation seen in the natural population. FIGG is designed to use high performance computing to rapidly generate artificial genomes, and can be used to generate large numbers of similar whole genome sequences by iteratively seeding each run with new parent genomes.

In the last few years high-throughput sequencing (HTS) has allowed researchers to sequence genomes for species that range from bacteria and plants, to insects and vertebrates. In the context of biomedicine HTS is being used to: characterize complex ecologies such as the human gut microbiome [1]; understand parasitic diseases such as malaria [2]; identify genomic variations that may be responsible for virulence in diseases such as tuberculosis [3]; and search for the mutations that drive genomic diseases such as cancer [4-6].

A result of this wide-ranging use of sequence information is petabytes worth of genomic data across multiple species, populations and diseases. New tools are constantly being required to enable the management and analysis of this information. The FIGG tool is meant to be of use to different computational researchers working in the area of large-scale genomics. In particular it is designed to be used by those who are struggling to keep pace with the scale and diversity of data in large-scale genomic projects. Using FIGG to generate artificial data has a number of advantages over downloading and storing publically available whole genome sequences as it:

* Correspondence: antonio.delsol@uni.lu
Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, 7, avenue des Hauts fourneaux, Esch/Alzette L-4362, Luxembourg

has known characteristics, so can be used for consistent benchmarking; can be used to generate mixed populations of heterogeneous genomes for algorithm testing; has no security requirements, so can be shared and used more easily; and does not place undue load on local resources, as genomes can be generated on the fly.

FIGG is designed to generate large volumes of potentially related sequences that can be used by computational researchers in testing their models, analysis pipelines and informatics solutions. Simulating experimental data is a common step in the development and evaluation of new analysis tools [7], computational methods, and the support infrastructure for managing such sequences. Many different genomic simulators are available (see Table 1) and have been described elsewhere [8], however these are not designed to provide the high volumes of complete genome sequences which are required for software testing and algorithm development. They range in application from instrument-specific sequence read simulation (e.g. ART [9], MetaSIM [10]), to genotype simulation for case–control studies based on linkage disequilibrium patterns (e.g. genomeSIMLA [11], GWASimulator [12]), to evaluating a population over time to determine how genomic hotspots or population bottlenecks affect a genome (e.g. FreGene [13], GENOME [14]) or protein sequence (e.g. ALF [15]).

FIGG generates whole genome sequence files, in FASTA format, by directly sampling from populations of observed variations. Each artificial genome includes sequence mutations that range from single nucleotide variations (SNV) to small and large-scale structural variations (e.g. indels, tandem duplications, inversions). It has been designed to use a distributed computing framework to enable rapid generation of large numbers of genomes while tracking the mutations that are applied to each. Below we provide details of the FIGG methods that enable the creation of diverse whole genomes which accurately model experimentally derived real sequence data. The following sections describe the methods used for analysis of background genomic variation, generation of the sequences, and validation of the models through the use of standard sequence analysis tools. Finally we discuss applications for FIGG within the sequencing community.

Methods

FIGG requires two inputs in order to create a genome: 1) all FASTA files representing the chromosomes to be simulated (e.g. chromosomes 1–22, X, and Y from human genome build GRCh37), and 2) a database that is the result of the frequency analysis as described in the next section (the full database format can be found at the link provided in Availability). The resulting output from FIGG is set of FASTA formatted sequence files (one per chromosome) that can be used by any tools which use FASTA as an input, including sequence-read simulators and genome alignment software.

Variation frequency analysis

The public availability of large datasets that characterize human genomic variability provide a wealth of data on population and individual variations. In order to develop an accurate estimate of the range of "normal" variation we used Ensembl [16]. This data was mined for all variants validated in the 1000Genomes [17] and HapMap [18] projects, as these are generally considered

Table 1 Genome simulators

Tool	Description	Outputs
ART [9]	Simulation of sequence reads with error models for multiple platforms (454, Solexa, SOLiD).	Single or pair ended sequence reads.
MetaSIM [10]	Simulation of sequence reads for metagenomics, particularly for highly variable data (taxonomically distinct but related organisms).	Single or pair ended sequence reads.
GENOME [14]	Population simulation within a set of alleles using genome level events such as recombination, migration, bottlenecks, and expansions.	Alleles identified as mutated (1) or not (0) across the simulated population.
GWASimulator [12]	Simulation of loci across a population which follows a given LD structure in case–control type studies.	SNVs per individual for input loci.
FreGene [13]	Mutation simulation using a theoretical sequence of a given size with hotspot, conversion, and selection parameters.	Mutation selection across population for a theoretical sequence.
genomeSIMLA [11]	Simulation of disease loci within a family or case–control setting using specific LD patterns for investigations of disease.	Affy identified SNPs selected by disease association.
ALF [15]	Population simulation for a specific gene set using a model for variation at the sequence and individual level.	FASTA protein and DNA sequences for specific genes.

Example simulators used in various types of genome investigations. Many use the Wright-Fisher model of population genetics theory [8] in order to generate populations that vary over time given some set of event frequencies such as LD, hotpots, population bottlenecks (GENOME, genomeSIMLA, FreGene), others provide a set of sequences that could be generated by a given sequencing technology with an error model (ART and MetaSIM). The specific simulator used is based on the type of investigation. In planning new GWAS studies for instance, a simulator that uses LD patterns and can provide predicted genomic regions for disease related mutations would be selected. However, such a simulator would not be of use in the planning of a metagenomic study for an organism which may not yet be fully sequenced, or is highly variable. None of these simulators provides whole genome FASTA as outputs.

representative of normal populations. Several other sources representing disease variations were downloaded for comparison, including those from the Catalogue of Somatic Mutations in Cancer (COSMIC) [19] and small structural variants in the Database of Genomic Variants Archive (DGVa) [20].

In order to characterize the variant frequency across the genome for different classes of mutations each chromosome was first fragmented into base-pair lengths that were manageable for processing. For each fragment a profile of unique variants was developed. These profiles were then analyzed to determine the frequency of each variant class: single point mutations being the most common, followed by sequence alterations (defined as an uncharacterized change in the sequence), and then insertions. Based on these frequencies structural elements in the sequence fragment were identified that can be directly observed and which could explain the variation frequencies including: a higher incidence of coding/non-coding regions; predicted CpG methylation sites; and high/low GC content. A weak correlation with SNVs was observed in segments with high/low GC content [21,22], but no other genome-wide structural correlation was found. When the same analysis on "disease" variations was run (e.g. COSMIC, DGVa) as a comparison, GC content continued to be the only clear structural correlation for variation frequency (see Figure 1 for a description of the final output).

Based on this analysis the observed sequence fragments were separated into bins by GC content, with variant counts per segment recorded for each chromosome (see Figure 2 for an example of the variant and GC tables in chromosome 4). The result is a set of tables that can be easily sampled for fragments based on a GC profile. Additionally, base pair size probabilities were calculated for all size-dependent variants (e.g. deletion sizes from 1–10 have a genome-wide frequency of 0.96, and from 11–100 a frequency of 0.04), and nucleotide mutation rates were determined for SNVs (e.g. C- > T 0.69, C- > A 0.16, C- > G 0.15, etc.).

Implementation

The general architecture of FIGG is shown in Figure 3. It has been designed to take advantage of distributed computing by both breaking down the processing of the data into a distributed model, and by separating the functionality required into distinct steps, called "jobs", that can be added or altered for downstream analysis or testing needs. FIGG is separated into three distinct jobs. The Additional file 1 document provided describes how to set up and run these jobs on an Amazon Web Services cluster.

The first job fragments a reference genome and persists it to a distributed database, which ensures that the background genomic information is highly accessible, and only needs to be run once per reference (e.g. GRCh37).

The second job mutates each of the segments from a parent genome, using information pulled from a variation frequency database. This database provides the information necessary to determine which variations should be applied to a given fragment (e.g. SNV, deletion, insertion) and how often these occur.

The third job assembles the mutated fragments into a whole genome, and generates the corresponding FASTA files. The second and third jobs are run in parallel to each other, allowing for a means to generate large numbers of artificial genomes in a highly scalable manner.

Mutation rules

The generation of new, mutated sequences is achieved through application of a ruleset based on the frequency analysis described above. Each input chromosome is split into fragments of the same size as those used for the frequency analysis (e.g. 1 kb). Each fragment is then processed stepwise (see Figure 4):

1. Determine the GC content of the fragment then fit to the identified bins in the frequency database based on the fragment chromosome. This provides a set of observed fragments to sample.
2. Randomly sample an observed fragment from the set of fragments that fit the GC bin. This fragment will include 0..*n* counts for each variation type (e.g. SNV, deletion, substitution, etc.).
3. Apply each variant type to the fragment sequentially (e.g. deletions first, tandem duplications last). This is achieved through sampling without replacement random sites within the fragment for each mutation, applying size-dependent or SNV probabilities for that mutation to the site, and repeating until all variants have been applied to the sequence.

The resulting fragment may vary significantly from, or be nearly identical to, the original sequence depending on the selected variant frequencies. Use of random site selection for applying the mutations ensures that no specific population bias (e.g. if the population that is used to generate the frequency data is overrepresented for a specific variant) is introduced into the bank of resulting sequences. The final FASTA sequence then provides a unique variation profile.

MapReduce for multiple genomes

Applying this process to the human genome to create a single genome is slow and inefficient on a single machine, even when each chromosome can be processed in parallel. In fact, a basic version of parallelization took more than 36 hours to produce a single genome. Producing banks of such genomes this way is therefore computationally limited. However, mutating the genome in independent fragments

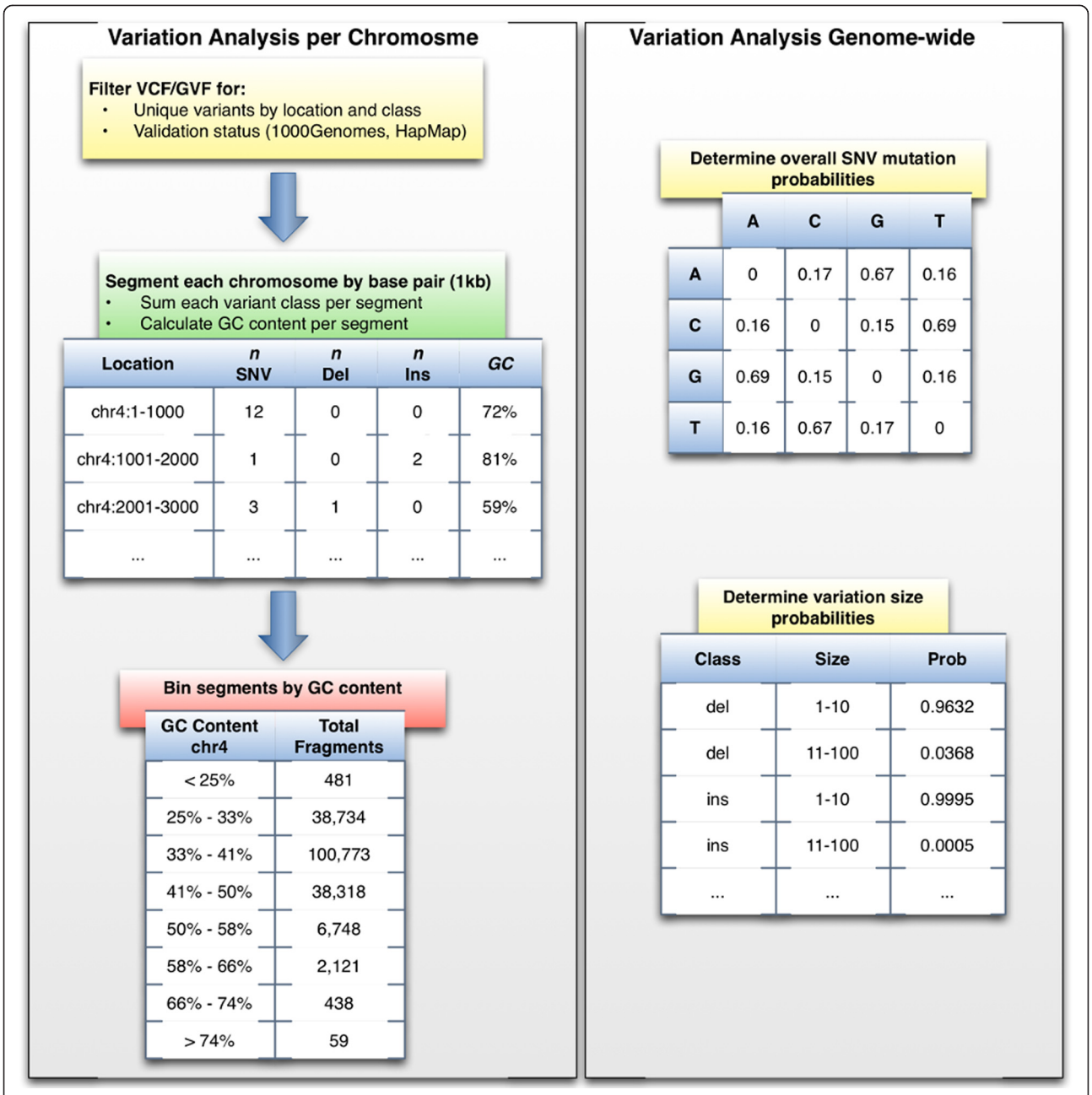


Figure 1 Variation frequency table generation procedure. The variation analysis uses publicly available small scale variation data to generate a set of database tables for a specific variation frequency. This is done in four separate steps. First, filter GVF or VCF files for unique variations per chromosome location and validation status. In this analysis variation files from Ensembl were used and "normal" validation status was determined based 1000Genomes or HapMap annotations. To generate a "highly variant" frequency, variations that were identified as being in the COSMIC and DGVa databases were added. Next, each chromosome is segmented into defined lengths (e.g. 1 kb) and the observed variations per class within the segment are counted. Additionally, the GC content for each segment is calculated from a corresponding FASTA sequence file. Then the segments are separated by GC content into 10 bins per chromosome. While these bins can be more granular, the correlation of SNV to GC content did not improve by increasing the number of bins. Finally, determine the genome-wide SNV mutation and size probabilities for variations that can be more than a single base pair in length. A database schema describing the final tables is provided in the source for FIGG.

makes this a good use case for highly distributed software frameworks such as Apache Hadoop MapReduce [23,24] backed by distributed file systems to create and store tens, hundreds, or more, of simulated genomes. In addition, use of HBase [25] allows for highly distributed column-based storage of generated sequences and mutations. This enables rapid scale-up for management, ensures that all variations to a given genome can be identified, and allows for the

Chromosome 4 - "Normal" Human Population

Fragments Per Bin

GC Content	Total Fragments
< 25%	481
25% - 33%	38,734
33% - 41%	100,773
41% - 50%	38,318
50% - 58%	6,748
58% - 66%	2,121
66% - 74%	438
> 74%	59

Variation Counts Per Fragment

Fragment (1..Total)	*n* SNV	*n* Del	*n* Ins	*n* SeqAlt
1	12	0	0	0
2	1	0	2	0
3	3	1	0	0
4	0	0	0	1
...	...	...	...	...
38,734	5	2	0	0

Figure 2 Variation frequency analysis. The result of the variation analysis is a table, indexed by chromosome and GC content, which provides experimentally observed counts of the different variations for that fragment. This means that a DNA fragment from chromosome 4 with a GC content of 25-35% has been observed 38,734 times. Each of those observed fragments is recorded with their variant counts. These observed fragments will be sampled from directly in the generation of an artificial genome.

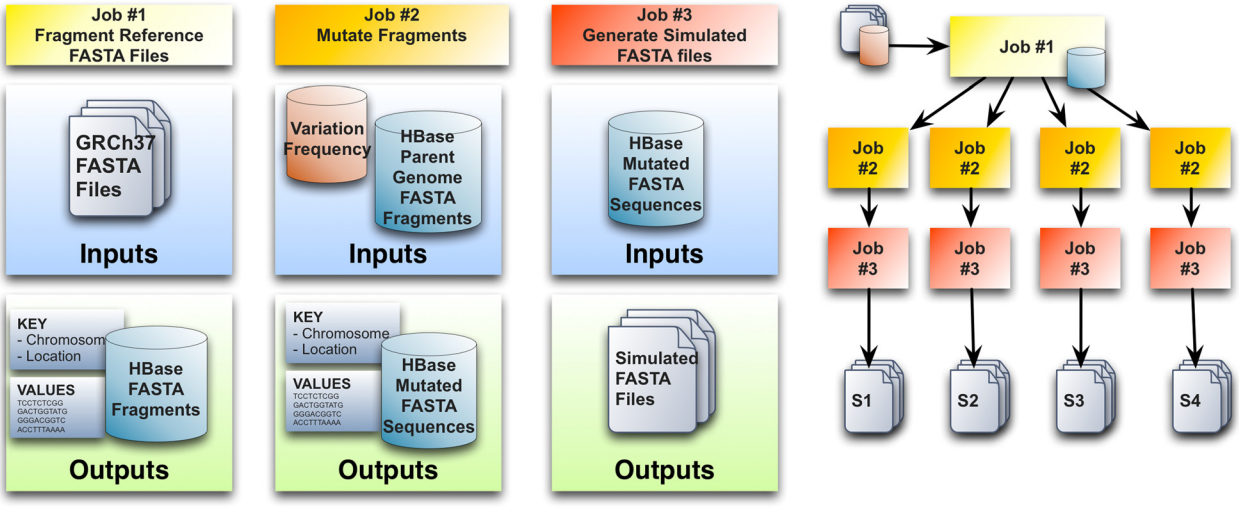


Figure 3 FIGG MapReduce jobs. Three discrete MapReduce jobs have been set up to generate unique whole genome sequences. The first job simply fragments the reference or "parent" genome into the distributed database, HBase. The second job reads all the fragments for the parent genome from the database, mutates them using the provided frequency information and again saves them to the database to ensure reproducibility. The final job generates FASTA formatted files, per chromosome, for the mutated genomes.

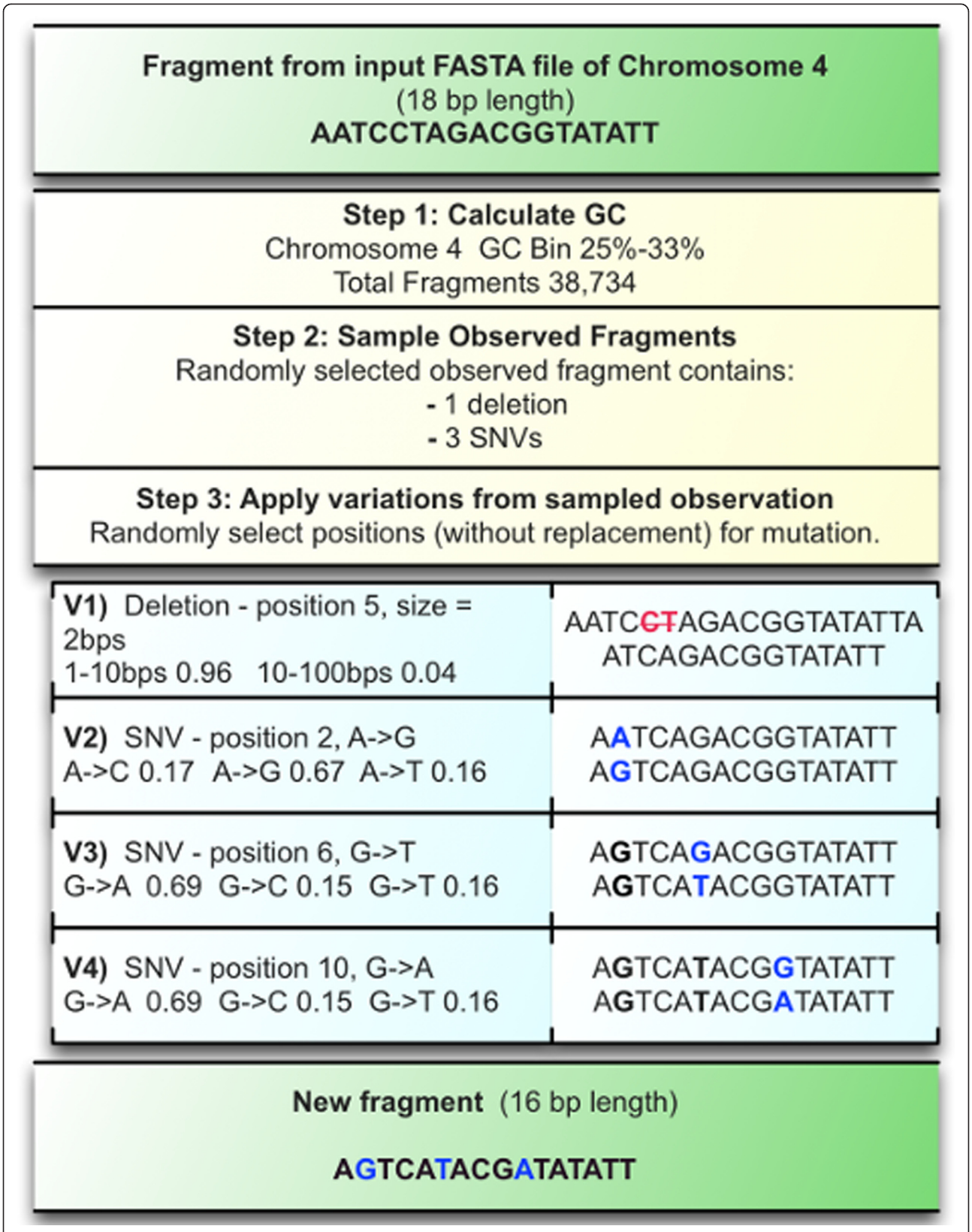


Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Fragment mutation rules. As an example of the process each fragment goes through, this fragment from chromosome 4 is mutated based on information from the tables shown in Figure 2. In step 1 the GC content of the fragment is calculated then fit to the pre-determined bins, all observed fragments within that bin are then available to sample. Step 2 samples one of these observed fragments to get the counts of specific variants. In this case the observed fragment had a single deletion and three SNVs. In step 3 these observed variant counts are applied in stages. Sites for each variation are selected randomly (without replacement), and the mutation applied. For a size-dependent variant such as the deletion, a size is determined from a probability table, for SNVs the probability of the point mutation is determined based on the nucleotide present at that site. The resulting fragment will not replicate the sampled fragment (from step 2) in specific mutations, but only in the number of mutations applied.

simple regeneration of simulated FASTA files on an as-needed basis.

MapReduce has been used effectively by us and others in various large-scale genomics toolsets to decrease computation times, and increase the scale of data that can be processed [26-28]. FIGG uses this framework in order to allow the rapid generation of new genomes or regeneration of previous mutation models. It is designed to run in three discrete jobs: 1) breakdown input FASTA files into fragments and save to a HBase database for use in subsequent jobs; 2) mutate all of the fragments from the first job and persist these to HBase; and 3) reassemble all mutated fragments as new FASTA formatted sequences.

MapReduce accomplishes these tasks by breaking each job into two separate computational phases (see Figure 5). The *Map* phase partitions data into discrete chunks and sends this to mappers, which process the data in parallel and emits key-value pairs. In each of the separate jobs for FIGG the mappers deal with FASTA sequences, either directly from a FASTA file or from HBase. Each mapper performs a computation on these sequences, and produces a sequence (the value) with a key that provides information about that sequence (e.g. chromosome location). These key-value pairs are "shuffle-sorted" and picked up by the *Reduce* phase. The framework guarantees that a single reducer will handle all values for a given key and that the values will be ordered.

It is worth noting that not all jobs will require the use of a reducer. In FIGG the first job which breaks down FASTA files into fragments and saves them to HBase (Job 1) is a "map-only" job, because we cannot further reduce these fragments without losing the data they represent. Therefore, the mappers output directly to HBase rather than to the reducers. In the mutation job (Job 2) the *Map* phase performs multiple tasks including applying variations to a sequence fragment, and writing new sequences and specific variation information directly to HBase. Whereas in Job 3 (FASTA file generation), the *Map* phase only does a single task, tagging a sequence with metadata that enables it to be ordered for the *Reduce* phase, which actually outputs the file. As each mapper is processing a subset of the data in parallel to all other mappers the compute time

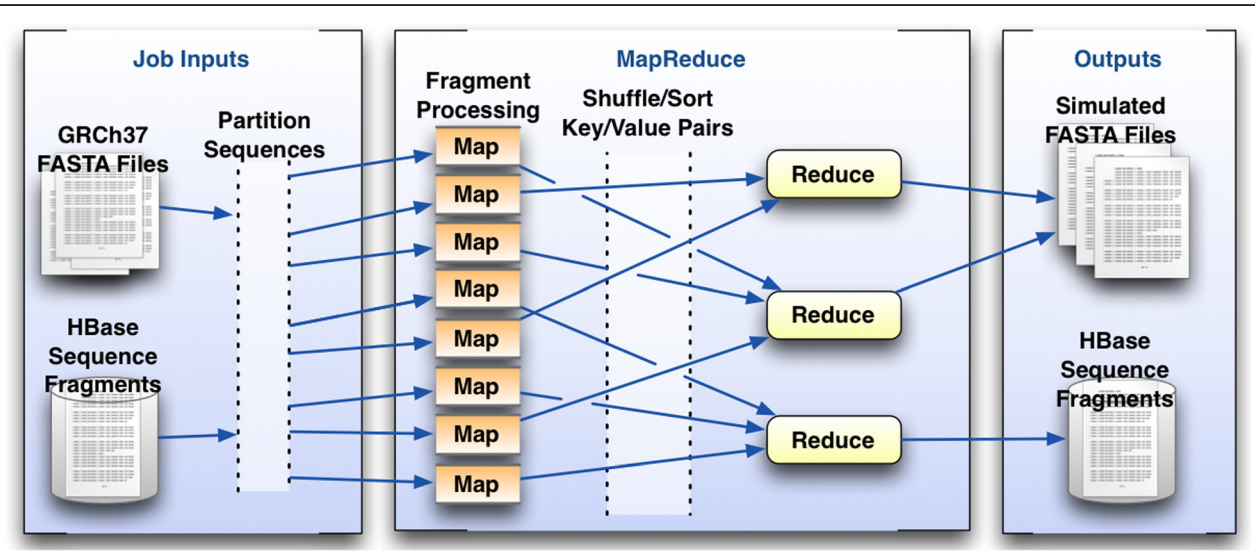


Figure 5 MapReduce framework. MapReduce provides a general framework to process partitionable data. The Map phase may either gather metadata statistics on a sequence fragment and write them to HBase (Job 1) or apply the variation frequencies and rules to a fragment (Job 2). The Reduce phase, if it is specified, is responsible for assembling the mutated fragments into FASTA formatted chromosome files (Job 3) or it may simply output additional metadata to HBase for use in other processing tasks.

required will scale directly with the number of mappers available to the task, limited in FIGGs case only to the organization of the data in HBase.

Results and discussion

Our primary interest in developing this tool was to provide sets of heterogeneous whole genomes in order to benchmark cancer genome alignments. This is a special case for alignment, as cancer genomes can vary quite dramatically between patients and even within a single tumor. With such a range of variation in patients, it was important to ensure that the simulated genomes were representative of the heterogeneity, without introducing biases for specific mutations.

In order to ensure that FIGG was modeling heterogeneous genomes that fit a specific background (e.g. "normal" or "diseased") two different frequency backgrounds were generated (see Methods). The "normal" frequency background was from data representative of the average human population: 1000Genomes and HapMap. The second, "highly variant" frequency background was based on data from the DGVa and COSMIC databases of cancer and other disease variations. This greatly increased the frequency and size of the small structural variations (e.g. millions of small deletions and insertions, up to several hundred bp in length).

Using these two different backgrounds and GRCh37 as the parent genome, FIGG generated six whole genome sequences: three "normal", two "highly variant", and one additional genome from the "normal" background that included a common cancer structural variation. As expected, for both the "normal" and the "highly variant" sequences, the simulated genomes preserved the frequency distribution of variations observed in the background data, while differing in the raw counts per fragment.

These simulated whole genomes were then used as references to align a set of low-coverage paired-end sequencing reads from the 1000Genomes project (NCBI Trace Archive accession ERX000272). The BWA alignment tool [29] was used to index the simulated genomes and align the reads against each reference, including the current reference genome GRCh37. Statistics regarding read mapping accuracy (see Table 2) for each genome were generated using SAMtools [30].

This comparison demonstrates that heterogeneous a whole genome sequences matching specific variation characteristics (e.g. normal, disease variant, etc.) can be generated by this tool. In the first three genomes the characteristics come from a "normal" population frequency and fairly closely match the mapping rates of the current public reference (GRCh37). The lower mapping rates in the high variation genomes are expected, as these will have a higher number of variations as well as longer insertions, deletions, and substitutions. This suggests that by using distributions for variations within distinct genomic populations, such as can be seen in different tumor types, highly specific simulated genomes can be generated. These specific simulated genomes could then be used as more accurate quality control sets for testing hypotheses or data. For instance, genome S6SV models a breakpoint that may be found in specific types of glioma [31-33]. This simulation could therefore be used to more accurately align a clinically derived sequence, integrate with proteomics data to infer a potential effect or biomarker, or simply provide a test sequence for breakpoint analysis methods [34].

Table 2 Sequence alignment statistics for simulated genomes

	SAMtools flagstat		
	Mapped	*Correctly paired*	*Singletons*
GRCh37	**98.22%**	**96.34%**	**0.85%**
S1	97.89%	95.52%	1.00%
S2	95.46%	92.95%	1.09%
S3	97.89%	95.54%	0.99%
S4H	90.09%	85.11%	2.89%
S5H	90.35%	85.45%	2.84%
S6SV	88.16%	83.22%	2.88%

A comparison of the 1000Genomes reads for ERX000272 mapped against each genome. GRCh37 is the current reference genome. S1, S2 and S3 are genomes generated based on normal variation data. S4H and S5H were generated with high variation data and S6SV is based on normal variations but with the chromosome arm 19q deleted. The table columns are statistics provided by SAMtools flagstat: *Mapped* provides the total percentage of reads that mapped to the genome on the left; *Correctly Paired* provides the percentage of reads that aligned to the genome in their proper pair; and *Singletons* provides the percentage of reads that were orphaned in the alignment. As expected, genomes S1-3 show mapping statistics that are close to the reference genome, while the others show a significantly lower statistics due to the higher frequency and larger bp size of variations used to generate these genomes.

Finally, it is important to note the benefits of using a highly distributed framework to generate these sequences. Current sequencing projects are generating hundreds or thousands of sequences from patients. In order to provide artificial data models to assist computational researchers working on large-scale projects, the simulation tool must be able to rapidly generate data of similar complexity and size. Distributed computing frameworks enable FIGG to generate this data quickly, allowing the researcher to simulate the scale of data they will actually be facing. Using Hadoop MapReduce enables FIGG to scale the mutation job nearly linearly to the number of cores available (see Figure 6). However, as with other distributed environments optimization for large clusters must be done on an individual basis.

Conclusions

HTS is now a primary tool for molecular biologists and biomedical investigations. Identifying how an individual varies from others within a population or how populations

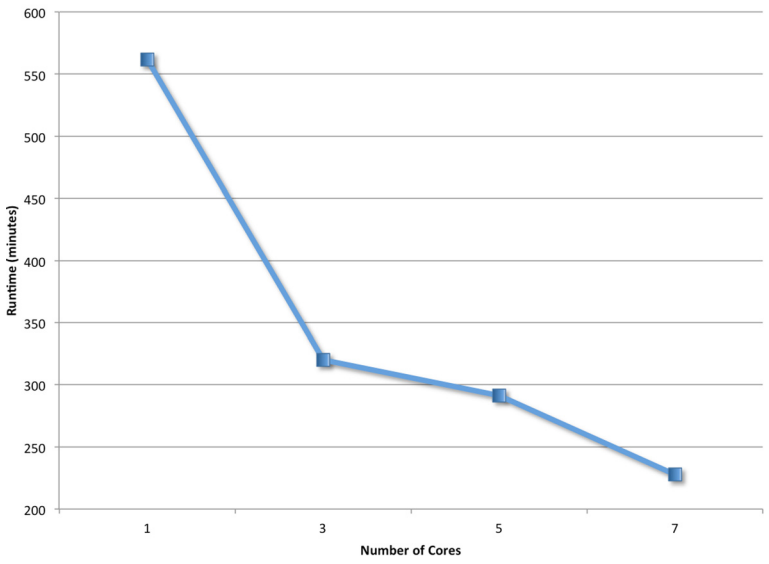


Figure 6 Scaling FIGG with MapReduce. The mutation process in FIGG is the most computationally intensive job in the pipeline. It was tested on Amazon Web Services Elastic MapReduce clusters of varying sizes for scalability. MapReduce provides a near linear speed up with the addition of nodes to this job. These genomes are saved to HBase to provide a persistent store of standard artificial genome data that can scale along with the cluster size. This is one area where optimization will provide increased performance as defining how the HBase tables are distributed can increase the speed of computation (e.g. more efficient row key design decreases query time and increases the number of available mappers). This is due to the fact that region server optimization is highly specific to the data, and improves as the data size increases.

vary from each other is central to understanding the molecular basis of a range of diseases from viral and parasitic, to autoimmune and cancer. As our understanding of these variations increases so too does the complexity of the analyses we need to undertake to find meaning in this data.

Simulation data is a common measure of the usability and accuracy of any analysis tools, but in whole genome studies there continues to be a lack of standard whole genome sequence data sets. This is especially problematic with the production of hundreds or thousands sequences from different populations. Comparing these to a single reference can lead to loss of important variation information found in even reasonably homogenous data. Highly heterogeneous populations, such as those found in cancer, may not even be represented at all by the reference. Generating thousands of whole genome models that vary predictably can provide highly specific test data for computational biologists investigating tumor diversity, software engineers who are tasked with supporting the large scale data that is being generated, and bioinformaticians who require reliable standards for developing new sequence analysis tools.

Central to each of these research needs is the development and use of banks of whole genome simulation data which will allow for the development of quality control tools, standard experimental design procedures, and disease specific algorithm research. FIGG provides simulation data models based on observed population information, will enable disease sequence modeling, is designed for large-scale distributed computing, and can rapidly scale up to generate tens, hundreds, or thousands of genomes.

Availability and requirements

Project name: Fragment-based Insilico Genome Generator
Home page: http://insilicogenome.sourceforge.net
Operating systems: Platform independent
Language: Java
Other requirements: Java version 1.6 or higher, A computational cluster running Hadoop v1.0.3 and HBase 0.92 (Amazon Web Services AMI v2.4.2), pre-computed HBase tables for the frequency analysis, and FASTA files for a reference genome.
Open source license: Apache 2.0
Restrictions for use: None

All Hadoop MapReduce jobs for this paper were run using Amazon Web Services MapReduce clusters. Please see the Additional file 1 for a walkthrough of the AWS job creation.

Additional file

Additional file 1: Amazon Web Services FIGG Walkthrough.

Abbreviations

COSMIC: Catalogue of Somatic Mutations in Cancer; DGVa: Database of genomic variants archive; HTS: High-throughput sequencing; SNV: Single nucleotide variation.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SK and AS conceived of, and planned project. SK analyzed variation data, implemented software and validated results. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by a grant from the Fonds National de la Recherche (FNR), Luxembourg [4717849] and Amazon Web Services Education & Research.

Received: 30 July 2013 Accepted: 9 May 2014
Published: 19 May 2014

References
1. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355–1359.
2. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, *et al*: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419:**498–511.
3. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, *et al*: **Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.** *Nature* 1998, **393:**537–544.
4. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PAW, Bignell GR, Stratton MR, Futreal PA: **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nat Genet* 2008, **40:**722–729.
5. Ley TJ, Mardis ER, Ding L, Fulton B, Mclellan MD, Chen K, Dooling D, Dunford-shore BH, Mcgrath S, Cook L, Abbott R, Larson DE, Koboldt DC, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Fulton L, Magrini V, Wylie T, Glasscock J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, *et al*: **DNA sequencing of a cytogenetically normal acute myeloid leukemia genome.** *Nature* 2008, **456:**66–72.
6. Atlas TCG: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061–1068.
7. Shrestha AMS, Frith MC: **An approximate Bayesian approach for mapping paired-end DNA reads to a reference genome.** *Bioinformatics* 2013, **29:**965–972.
8. Hoban S, Bertorelle G, Gaggiotti OE: **Computer simulations: tools for population and evolutionary genetics.** *Nat Rev Genet* 2011, **13:**110–122.
9. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator.** *Bioinformatics* 2012, **28:**593–594.
10. Richter DC, Ott F, Auch AF, Schmid R, Huson DH: **MetaSim: a sequencing simulator for genomics and metagenomics.** *PLoS ONE* 2008, **3:**e3373.
11. Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, Martin E, Ritchie MD: **Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA.** In *Evol. Comput. Mach. Learn. Data Min. Bioinforma. Lect. Notes Comput. Sci.* 4973rd edition. Berlin Heidelberg: Springer; 2008:24–35.
12. Li C, Li M: **GWAsimulator: a rapid whole-genome simulation program.** *Bioinformatics* 2008, **24:**140–142.
13. Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ: **Sequence-level population simulations over large genomic regions.** *Genetics* 2007, **177:**1725–1731.
14. Liang L, Zöllner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator.** *Bioinformatics* 2007, **23:**1565–1567.
15. Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C: **ALF–a simulation framework for genome evolution.** *Mol Biol Evol* 2012, **29:**1115–1123.
16. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM, Brent S, Kulesha E, Marin-Garcia P, Smedley D, Birney E, Flicek P: **Ensembl variation resources.** *BMC Genomics* 2010, **11:**293.
17. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061–1073.
18. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789–796.
19. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R: **The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.** *Br J Cancer* 2004, **91:**355–358.
20. **Database of Genomic Variation Archive.** http://www.ebi.ac.uk/dgva/.
21. Kudla G, Helwak A, Lipinski L: **Gene conversion and GC-content evolution in mammalian Hsp70.** *Mol Biol Evol* 2004, **21:**1438–1444.
22. Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** *Trends Genet* 2002, **18:**337–340.
23. Dean J, Ghemawat S: **MapReduce: Simplified Data Processing on Large Clusters.** *Commun ACM* 2008, **51:**1–13.
24. **Apache Hadoop.** http://hadoop.apache.org/.
25. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE: **Bigtable: A Distributed Storage System for Structured Data.** *ACM Trans Comput Syst* 2008, **26**(4):1–4:26.
26. Robinson T, Killcoyne S, Bressler R, Boyle J: **SAMQA: error classification and validation of high-throughput sequenced read data.** *BMC Genomics* 2011, **12:**419.
27. Schatz MC: **CloudBurst: highly sensitive read mapping with MapReduce.** *Bioinformatics* 2009, **25:**1363–1369.
28. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL: **Searching for SNPs with cloud computing.** *Genome Biol* 2009, **10:**R134.
29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754–1760.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078–2079.
31. Smith JS, Perry A, Borell TJ, Lee HK, O'Fallon J, Hosek SM, Kimmel D, Yates A, Burger PC, Scheithauer BW, Jenkins RB: **Alterations of chromosome arms 1p and 19q as predictors of survival in oligodendrogliomas, astrocytomas, and mixed oligoastrocytomas.** *J Clin Oncol* 2000, **18:**636–645.
32. Ręcławowicz D, Stempniewicz M, Biernat W, Limon J, Słoniewski P: **Loss of genetic material within 1p and 19q chromosomal arms in low grade gliomas of central nervous system.** *Folia Neuropathol* 2013, **51:**26–32.
33. **Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.** http://cgap.nci.nih.gov/Chromosomes/Mitelman.
34. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing.** *Nat Methods* 2009, **6:**S13–S20.

doi:10.1186/1471-2105-15-149
Cite this article as: Killcoyne and del Sol: FIGG: Simulating populations of whole genome sequences for heterogeneous data analyses. *BMC Bioinformatics* 2014 **15**:149.