ACCURATE MUTATION DETECTION IN CANCER THROUGH IMPROVED ERROR SUPPRESSION AND QUANTIFICATION OF MOLECULAR COMPLEXITY

by

Ting Ting Wang

A thesis submitted in conformity with the requirements for the degree of Master of Science
Department of Medical Biophysics
University of Toronto

© Copyright 2018 by Ting Ting Wang
Abstract

Accurate Mutation Detection in Cancer Through Improved Error Suppression and Quantification of Molecular Complexity

Ting Ting Wang
Master of Science
Department of Medical Biophysics
University of Toronto
2018

DNA from cancer cells is often masked by nucleic acids from normal cells in tumour tissues and blood plasma, resulting in low allele frequency mutations. Next-generation sequencing has enabled detection of somatic alterations, however it is limited by sequence errors. A promising strategy to suppress errors utilizes unique molecular identifiers (UMIs) to amalgamate reads from the same molecule into a consensus sequence. While current methods depend on redundant sequencing of template molecules, we have developed an error suppression strategy that only requires single reads. We expanded UMIs with sequence properties to enhance molecular quantification, which reduced single-strand consensus errors up to 70% for deep sequencing. Additionally, we suppressed errors down to $1 \times 10^{-4}\%$ in a quarter of single reads with moderate sequencing coverage. Using complementary strands of single reads, we improved recovery of double-strand molecules by 5-fold. Together, our method enabled sensitive mutation detection below 0.1% frequency.
Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Trevor Pugh, for his mentorship and insightful expertise over these past two years. He has provided me with an abundance of opportunities and exposure to clinical and translational research. His impeccable vision, optimism, and collaborative nature are qualities to which I aspire. I would like to extend my sincerest gratitude to my co-supervisor, Dr. Scott Bratman, who has been an integral component to this body of work and as well to my own personal development. I am grateful for his kindness and support through the toughest of times. His extensive knowledge in this field has been central to the progression of my research. I have been privileged to learn from two outstanding scientists and I would like to thank them for their guidance throughout my graduate journey.

I would also like to thank members of my supervisory committee, Dr. Mathieu Lupien and Dr. Sean Cleary, for their scientific support, feedback, and insightful career discussions. Thank you to Dr. Sagi Abelson for technical suggestions and aid in the development of the method presented in this body of work. I would also like to thank members of the Princess Margaret Genomics Centre (Nicholas Khuu, Julissa Tsao, and Zhibin Lu) for their assistance with sequencing and raw data processing. I am extremely grateful to members of the Pugh lab and the Bratman lab, both past and present. Thank you to Dr. Jeff Bruce, Arnavaz Danesh, Marco DiGrappa, and Rene Quevedo for their constant computational support, as well as Zhen Zhao, Dr. Tiantian Li, and Iulia Cirlan for sample preparation and wet lab expertise. I deeply appreciate the companionship of the truly amazing people I am surrounded by daily, as they have made each and every day a delight.

Finally and most importantly, I want to thank my family and friends for accompanying me through my graduate pursuit. To mom and dad, thank you for your never-ending encouragement and unconditional love and support for everything that I do. To my friends, thank you for challenging me to embrace new experiences and inspiring me with your compassion. In particular, I want to thank Ingrid Kao and Parasvi Patel for building with me life-long friendships and Tommy Hu for always bringing a smile to my face. I am extremely grateful to have shared so many memorable moments with all of you. Thank you.
1 Introduction

1.1 Next-generation sequencing in oncology

1.1.1 Challenges with tissue biopsy

1.1.2 Overview of liquid biopsy

1.1.3 Origins and characteristics of circulating tumour DNA

1.2 Clinical implications of circulating tumour DNA

1.2.1 Earlier diagnosis of disease

1.2.2 Minimal residual disease and recurrence monitoring

1.2.3 Evaluating patient response to treatment

1.3 Challenges with low-allele-frequency variant detection

1.3.1 Pre-analytical barriers

1.3.2 Technology platforms for rare variant analysis

1.3.3 Library complexity and artefactual errors

1.3.4 Molecular quantification and error suppression strategies

1.4 Rationale and hypothesis

1.5 Objectives

2 Materials and Methods

2.1 Target panel design

2.2 Next-generation sequencing library preparation

2.3 Target capture and sequencing

2.4 Data preprocessing

2.5 Molecular barcode analysis

2.5.1 Unique molecular identifiers

2.5.2 Single strand consensus sequences

2.5.3 Singleton rescue (SR)
2.5.4 Duplex Consensus Sequences .................................................. 25
2.6 Evaluation of consensus sequence formation ................................. 25
2.7 Error analysis .............................................................................. 26
2.8 in silico downsampling ............................................................... 26
2.9 Mutation analysis ................................................................. 27

3 Results .................................................................................. 28
3.1 Sequence properties can enhance barcode depiction of molecular diversity 28
3.2 Singleton rescue improves efficiency of consensus formation ........... 30
3.3 Singleton rescue achieves comparable rates of error suppression as duplex correction .................................................. 32
3.4 Oxidative damage is effectively suppressed by singleton rescue ........ 34
3.5 Duplex molecular recovery is maximized through singleton rescue .... 35
3.6 Singleton enabled error suppression improves lower limit of mutation detection .................................................. 39

4 Discussion ........................................................................ 46
4.1 Conclusions ........................................................................... 46
4.2 Potential impact and applications ........................................... 47
4.3 Future directions ................................................................. 48

Bibliography ........................................................................ 50
List of Tables

2.1 Summary of library preparation specifications for each cell line dilution. . 18
List of Figures

1.1 Sources and alterations of cell-free DNA. cfDNA can be derived from a multitude of cells and tissue types: tumour, stroma of the tumour microenvironment, immune, and other tissues or organs. Cells can release cfDNA through apoptosis, necrosis, and secretion. Genetic and epigenetic alterations can be used to infer tumour-derived cell-free DNA. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [1], 2017.

1.2 ctDNA applications for disease management. Clinical applications of ctDNA analysis in a hypothetical patient throughout course of disease. ctDNA could enable earlier diagnosis and molecular profiling of tumours prior to metastatic spread. Localized cancer is more amenable to surgery and treatments with curative intent. Detection of minimal residual disease post-operation could inform of relapse and neo-adjuvant therapy can be monitored with serial liquid biopsies to elucidate clonal evolution. This schematic illustrates the development of a single tumour clone that diverges into multiple distinct entities as a result of selective pressures from cancer progression and drug therapies. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [1], 2017.

1.3 Stages impacting molecular complexity and artefactual errors. a) Detection of ctDNA is impacted by molecular complexity and is limited by level of background errors. While adapter ligation and target capture may impact molecular complexity, the lower limit of detection is determined by the rate of artefactual errors. These sequence mistakes may arise from polymerase errors, sequencing blunders, or oxidative damage during sample extraction and target capture. b) Raw sequence reads are not representative of true molecular diversity due to PCR duplicates. De-duplication using genome coordinates is not an accurate depiction either as independent molecules may map to the same position. UMIs enable more accurate quantification of molecular complexity and can enable sequence error suppression by correcting artefacts through consensus formation. Although traditional barcoding methods rely on duplicate reads, we have developed a strategy to enable error suppression in single reads.
2.1 UMIs: A barcode family is defined by barcodes and sequence features that enable unique identification of an individual molecule. Our 1) 4-bp barcode allows for 256 different molecules at each 2) genomic position. In addition, 3) CIGAR strings can provide additional identifier diversity as it encodes sequence variation to the reference genome highlighting distinct attributes such as clipped regions, insertions, or deletions. Furthermore, read 4) orientation and 5) number can differentiate palindromic barcodes and provide proper molecular grouping of reads belonging to different strands. R1 = Read 1 and R2 = Read 2 of paired-end reads.

2.2 Source of CIGAR: When DNA sequencing reads (grey bars) are aligned to a reference sequence, specific genome coordinates and mismatches to the reference (CIGAR strings) are noted in the sequence alignment file. CIGAR strings encode information about fragment clipping (S, soft clipped bases in 5’ or 3’ end of the read that are not part of the alignment), as well as sequence concordance with the reference genome (M, match/mismatch bases), insertions (I), and deletions (D).

2.3 Duplex sequencing schematic: An uncollapsed BAM file is first processed through SSCS_maker.py to create an error suppressed single strand consensus sequences (SSCS) BAM file and an uncorrected Singleton BAM file. The single reads can be recovered through singleton_strand_rescue.py, which salvages singletons with its complementary SSCS or singleton. SSCS reads can be directly made into duplex consensus sequences (DCS) or merged with rescued singletons to create an expanded pool of DCS reads (Figure illustrates singleton rescue merged work flow).

3.1 Impact of alignment information on molecular complexity and error suppression. a) Fraction of reads with alternative alignment to the reference genome in SSCS and singletons across two cell line dilutions. Error bars represent standard error between samples. b) CIGAR strings are indicative of alignment as they encapsulate information about base concordance, fragment clipping, insertions, and deletions. Scatter plots show change in singletons and SSCS error rate with CIGAR embedded identifiers compared to reads derived without alignment differentiation. Left: a large panel sequenced to moderate depth (LgMid, n = 12). Right: a small panel with deep sequencing (SmDeep, n = 8).
3.2 Dual strategies of single read error suppression: Singleton rescue by SSCS starts with collapsing duplicate reads to form SSCS, which are subsequently used to correct the reverse complement singleton. Additionally, singletons can be rescued by other singletons through complementary strand error suppression.

3.3 Proportion of reads with duplicate support eligible for consensus-based error suppression. Singleton rescue (SR) enables additional error correction of singletons through rescue strategies using complementary SSCS or singletons. The left shows the average read distribution of a large panel with sequencing to moderate depth (LgMid, n = 12), while the right shows a summary of reads from a small panel with deep sequencing (SmDeep, n = 8).

3.4 Error suppression performance at each stage of error correction between sequencing with moderate (LgMid) and deep (SmDeep) coverage. Background error determined by non-reference bases at allele frequencies below 5%. Error bars represents standard error across all samples.

3.5 Duplex strategies of error correction suppresses oxidative damage. a) Schematic of oxidative damaging induced 8-oxoguanine (8-oxoG) asymmetric strand lesions in DNA. 8-oxoG arises from oxidation and converts to thymine after PCR amplification. As only plus strands are captured with our target panel comprised of negative probes, there is a bias for C>A errors. b) Comparison of 12 substitution classes between uncorrected and error suppressed sequences in LgMid (n=12). c-g) Ratio of errors between reciprocal base substitutions. Imbalances between C>A/G>T is indicative of oxidative damage.
3.6 Singleton rescue impact on consensus formation across a wide coverage range (500x to 128,000x). Results are shown for LgMid and SmDeep cell line dilutions that were down-sampled by half at each interval step. Error bars represents standard error of all the samples across ten simulations. Mean target coverage is log10 transformed to show trends at lower range. **a)** Proportion of singletons rescued through complementary SSCS and singleton correction. **b)** Efficiency of SSCS formation as determined by number of SSCS / total uncollapsed reads. **c)** DCS efficiency as calculated by number of DCS / total uncollapsed reads **d)** DCS recovery rate of observed / expected DCS. Expected DCS corresponds to half of all SSCS reads as two single molecules should theoretically form a DCS. Efficiency is an assessment of over-sequencing relative to unique input molecules, whereas recovery is an estimate of molecular retrieval.

3.7 SmDeep: three HCT116 mutations identified through deep sequencing of a 5 gene panel across dilutions from 100% to 10^{-4}%. Every base at each position was called to illustrate the frequency of background errors present. Top panel compares limit of detection with standard uncollapsed reads, middle panel represents single strand consensus sequences, and bottom panel highlights duplex consensus sequences.

3.8 LgMid: Five MOLM13 mutations identified through sequencing to moderate depth across dilutions from 100% to 0.04%. Every base at each position was called to determine frequency of known mutations and as well to illustrate the frequency of background errors. Each row corresponds to an error suppression method and each column represents a different mutation.

3.9 Summary trends of SmDeep mutations across different error suppression strategies. Comparison of detection limit using **a)** allele frequency and **b)** reads supporting each variant.

3.10 Summary trends of LgMid mutations across different error suppression strategies. Comparison of detection limit using **a)** allele frequency and **b)** reads supporting each variant.
List of Terms and Abbreviations

8-oxoG  8-oxoguanine lesions
bp     Base pairs
CCLE   Cancer Cell Line Encyclopedia
CIGAR  Concise Idiosyncratic Gapped Alignment Report
tDNA   Circulating tumour DNA
cfDNA  Cell-free DNA
CTC    Circulating tumour cell
CDX    CTC-derived xenograft
DCS    Duplex / double-strand consensus sequence
DNA    Deoxyribonucleic acid
dPCR   Digital PCR
EDTA   Ethylenediaminetetraacetic acid
LLOD   Lower limit of detection
MAPK   Mitogen-activated protein kinase
miRNA  MicroRNA
MRD    Minimal residual disease
NGS    Next-generation sequencing
NSCLC  Non-small cell lung cancer
PCR    Polymerase chain reaction
qPCR   Quantitative PCR
SR     Singleton rescue
SSCS   Single strand consensus sequence
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TKI</td>
<td>Tyrosine kinase inhibitor</td>
</tr>
<tr>
<td>UMI</td>
<td>Unique molecular identifier</td>
</tr>
<tr>
<td>WBC</td>
<td>White blood cells</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
</tbody>
</table>
Glossary

Allele frequency  Frequency of an allele at a particular locus or position.

CIGAR  Sequence alignment information relative to the reference genome.

Duplex consensus sequence  Consensus sequence derived from complementary single-strand consensus sequences of a molecule.

Mean target coverage  Average number of reads per position within a target region.

Molecular diversity  Unique DNA molecules within a sample, inferred by bioinformatic removal of PCR duplicates.

Molecular recovery  Proportion of original inferred molecular diversity retained after the process of library preparation, sequencing, and bioinformatic analysis.

Molecular saturation  The point of sequencing beyond which additional sequencing does not result in increased molecular diversity or recovery.

Read  String of consecutive bases from a single molecule of DNA produced by next-generation sequencing.

Singleton  Molecules that have been sequenced only once (i.e., without PCR duplicates in the sequence data).

Singleton rescue  Single read error correction through consensus formation with the complementary strand.

Single strand consensus sequence  Consensus of all reads derived from the same strand of a template DNA molecule.

Uncollapsed reads  Mapped reads that have not been subjected to de-duplication, error suppression, or other analysis related to molecular identity using UMIs or barcodes.

Unique molecular identifier  A series of bases or barcodes added to a molecule either through ligation or amplification. After sequencing, these barcodes can be used to identify which original template DNA molecule a particular read is derived from.
Chapter 1

Introduction

1.1 Next-generation sequencing in oncology

Cancer arises through an accumulation of molecular alterations affecting cellular growth and survival. Cells become malignant as they undergo continuous proliferation, evasion of growth suppressors and cell death, induction of angiogenesis, and initiation of tumour invasion and metastasis [2]. This transformation is the result of combined destabilization of many important cellular processes. As cancer evolves dynamically with development of disease, tumours become heterogeneous harbouring a diverse array of cells with distinct molecular profiles. Pressures applied by treatment can select for cancer cells with specific genome alterations that confer a growth advantage. Despite acquiring similar advantageous phenotypes, these alterations can differ between tumours, highlighting the need for comprehensive molecular profiling, including massively parallel next-generation sequencing of tumour DNA[3].

Diverse molecular alterations can propagate within a single tumour, between multiple tumour sites within a patient, or across multiple patients. Inter-tumour heterogeneity is the variability between patients with distinct molecular and histological profiles. Patients can be classified into subgroups based on these different features to improve prediction of prognosis and response to therapy. Intra-tumour heterogeneity refers to the variability between tumour cells of an individual. This can manifest as spatial diversity across different disease sites or within a tumour, as well as temporal variations in the molecular composition of cancer cells over time [4, 5]. Selective pressure from treatment may confer resistance through expansion of subclonal populations or evolution of drug tolerant cells. While heterogeneity allows tumours to overcome evolutionary pressures, it can also be exploited for therapeutic targets. Characterizing the molecular landscape of these mixed cell populations is essential in guiding precision medicine to improve patient response to therapy.

Next generation sequencing (NGS) has revolutionized the understanding of genetics by enabling broader characterization of the genome. NGS has had profound implications in cancer by uncovering alterations responsible for the development of disease. Single nucleotide and structural variations with high allele frequencies can be profiled using NGS.
for molecular characterization of tumour heterogeneity. However, detection of subclonal mutations at lower frequencies (<1-2%) has been limited by technical artefacts and sequencer errors [6]. Whole genome sequencing (WGS) identifies a range of alterations including copy number aberrations, nucleotide substitutions, and structural rearrangements. This can enable discovery of novel chromosomal rearrangements and somatic mutations of non-coding regions that were previously unobserved with other methods.

While WGS provides the most comprehensive profiling of the genome, it is burdened by the high cost associated with the massive amount of sequencing it requires. Typically, 90 Gb of sequence data is needed to cover 3 billion bases of the human genome with an average of 30x coverage per base [7]. As only 1% of the human genome is comprised of coding exons of genes, whole exome sequencing can achieve 100-fold greater coverage (3,000X) using the same number of reads. Targeted sequencing allows greater coverage across regions of interest at a lower cost. Under the same scenario, a target panel of 0.3 MB can be sequenced 300,000x per position with the same number of reads as WGS. Thus, deep molecular profiling with targeted sequencing has the potential to enable detection of low-allele-frequency mutations despite admixtures of tumour and non-malignant cells.

1.1.1 Challenges with tissue biopsy

Heterogeneity confounds identification of biomarkers indicative of response, which is critical for patient selection and the success of targeted therapies. At the present time, patient profiling is primarily achieved using a single tissue resection or biopsy sample. As an individual tumour sample is not representative of diversity within primary or metastatic lesions, this greatly confounds patient stratification for personalized medicine. In addition, somatic mutations can be obscured by non-malignant cells within the tissue. Clonal populations in one biopsy may be subclonal or absent in subsequent tumour sampling. These invasive sampling procedures are subject to complications and are bound by the accessibility of the tumour [8]. Moreover, the impracticality of serial tissue biopsies preclude the possibility for multi-site genotyping or longitudinal profiling. There is a clinical need for new molecular and diagnostic tools as current methodologies only provide a limited scope on the complexity of cancer.

1.1.2 Overview of liquid biopsy

Liquid biopsy has the potential to transform cancer management by overcoming barriers inherent to conventional biomarkers [9]. Cells, exosomes, and nucleic acids found in the
blood or other bodily fluids enable a non-invasive holistic view of cancer clonal populations [10, 11]. Circulating tumour cells (CTCs) shed from primary or metastatic lesions can be enriched through size selection or surface markers. These cells can then be expanded and characterized to facilitate the development of personalized cancer medicine. Although challenges exist with culturing CTC cell lines, CTC-derived xenograft (CDX) models have been shown to mirror donor response to chemotherapy [12]. As CDXs share similar genomic profiles to their donors, new drug targets could potentially be identified by comparing models derived before and after treatment resistance. However, CTCs are found at a frequency of approximately 1 tumour cell per billion blood cells in metastatic cancer patients. While the low frequency and the time required to establish in vivo models impede its immediate clinical utility, CTCs still have significant implications for research [13].

In addition to intact cells, extracellular vesicles important for intercellular communication can be isolated through centrifugation or marker selection. Their RNA contents can be sequenced for cancer-specific gene expression profiles [14]. This approach may be useful for patients with limited quantities of detectable circulating tumour DNA, as mRNA from highly expressed genes are of greater abundance relative to the two copies present in circulating tumour DNA. While cell-free RNA can also be detected in the circulation, mRNA are of low abundance due to plasma nucleases. MicroRNA (miRNA) are more stable as they are protected by the argonaute-2 protein complex or encapsulated in extracellular vesicles [15]. As miRNA transcripts differ between cancer and healthy individuals, they may be potential candidates for disease biomarkers [16]. Although CTCs and exosomes provide insight on the molecular architecture of cancer populations, their utility is hindered by their limited availability and poor scalability to be translated into the clinic [17]. Similarly, applications of cell-free RNA are also controversial as it is unclear whether they are of tumour origins [9]. For these reasons, among the various potential analyses for liquid biopsy applications, circulating DNA may hold the greatest promise for providing clinical impact in oncology.

1.1.3 Origins and characteristics of circulating tumour DNA

Although the presence of cell-free nucleic acid was first described more than half a century ago [18], observations in oncology were not reported until decades later and its implications for cancer patients were not extensively explored until recently [19]. With the advent of next-generation sequencing, genetic and epigenetic alterations in the tumour can be identified in the mutant DNA fragments of the peripheral blood. Aside from the
Cell-free DNA (cfDNA) can be derived from a multitude of cells and tissue types: tumour, stroma of the tumour microenvironment, immune, and other tissues or organs. Cells can release cfDNA through apoptosis, necrosis, and secretion. Genetic and epigenetic alterations can be used to infer tumour-derived cell-free DNA. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [1], 2017.

blood, circulating tumour DNA (ctDNA) is also evident in the cerebrospinal fluid for central nervous system cancers and pleural effusion fluids for respiratory system cancers. For truly non-invasive interrogation, ctDNA can be collected from the stool for colorectal cancers, urine for bladder cancers, and saliva for head and neck cancers [9]. Concentration of ctDNA may be higher in bodily fluids in greater proximity to the tumour, however detection in the blood has the broadest clinical applications across multiple cancers.

Cell-free DNA (cfDNA) can be derived from a multitude of cells and tissue types: tumour, stroma of the tumour microenvironment, immune, and other tissues or organs. While the mechanism of cfDNA origins is unknown, fragments are suggested to have been released into the circulation via apoptosis, necrosis, and secretion (Figure 1.1). cfDNA has a characteristic peak (166 bp) corresponding to the length of nuclease-cleaved DNA from apoptotic degradation. This length encompasses the span of DNA wrapped around histones (˜147 bp) and linked between nucleosome cores [20]. Using long-read sequencing technology, large cfDNA fragments (>1,000 bp) have also been identified and may be associated with exosomal or necrotic release [1]. Interestingly, cfDNA from non-mutant cells are longer than those that are tumour derived (˜166 bp vs. 132-145 bp)
This is also observed in animal xenografts, where tumour fragments can be easily distinguished by their human origin from background rat cfDNA. Differences in length between mutant and wild-type cell-free molecules is potentially caused by variability in nucleosome wrapping. This distinction between DNA length could be exploited for more efficient isolation of tumour fragments.

Genome-wide sequencing of cfDNA can be used to infer nucleosome occupancy or methylation profiles enabling determination of cell-types of origin [22, 23]. Germ line cfDNA from non-mutant cells represent the majority of cell-free molecules and they are primarily derived from haematopoeitic cells in healthy individuals [23]. Despite representing only a small fraction of total cfDNA, ctDNA abundance in the plasma is prognostic of disease progression. While ctDNA can be found ranging below 0.1% to above 10%, it has been shown to correlate with cancer stage and tumour size [24, 25]. Notably, the quantity of cell-free molecules can also increase with exercise, infection, pregnancy, or smoking [23, 9, 1]. Although cfDNA is continuously shed, their half-life ranges between 16 minutes and 2.5 hours with clearance from the circulation via kidneys, liver, spleen, and circulating nucleases [1]. This constant overturn of cfDNA enables more dynamic monitoring of patient response than other blood-based biomarkers [26].

In recent studies, ctDNA has demonstrated clinical utility with capabilities to recapitulate both spatial and temporal heterogeneity [9]. With continuous release and rapid clearance of circulating cfDNA, ctDNA can provide a real-time snapshot of tumour disease burden. ctDNA is shed by both primary and metastatic lesions and its abundance has been shown to correlate with cancer stage and tumour size [25]. In a study with 640 patients across various stages and cancer types (breast, colorectal, gastroesophageal, pancreatic), ctDNA was detected at a frequency of fewer than 10 copies per 5ml of plasma in early-stage patients. In contrast, ctDNA was found at a median concentration of 100-fold higher in advanced stage malignancies [24]. Analysis of ctDNA can provide earlier indications of relapse and outperform conventional blood-based assays [27, 26]. The presence of these tumour-derived DNA fragments containing somatic mutations after therapy is indicative of recurrence and can be detected 5 months prior to radiographic imaging [26]. In turn, serial blood sampling can elucidate clonal evolution and permit dynamic longitudinal disease monitoring [28]. In summary, ctDNA has significant utility in oncology from earlier diagnosis to monitoring recurrent disease.
1.2 Clinical implications of circulating tumour DNA

Unlike tissue biopsies subject to complications and sampling biases, liquid biopsies provide a safer approach for comprehensive cancer profiling. The lower limit of detection (LLOD) for ctDNA is the threshold to which we can confidently discern mutations from background germ line cfDNA. Sensitivity of detection is measured by proportion of positive mutations correctly identified, whereas specificity measures the proportion of negatives correctly identified. ctDNA abundance is often influenced by disease burden and extent of tumour genomic characterization, which guides mutational search and governs scope of detection. With NGS technology, the LLOD of ctDNA analysis is dictated by the depth of sequencing as well as the level of technical artefacts from PCR or sequencing errors. This has significant implications in oncology where ctDNA abundance can be used to monitor patient response, while presence of ctDNA can be used for earlier diagnosis or identification of recurrent disease (Figure 1.2).

Figure 1.2: ctDNA applications for disease management. Clinical applications of ctDNA analysis in a hypothetical patient throughout course of disease. ctDNA could enable earlier diagnosis and molecular profiling of tumours prior to metastatic spread. Localized cancer is more amenable to surgery and treatments with curative intent. Detection of minimal residual disease post-operation could inform of relapse and neo-adjuvant therapy can be monitored with serial liquid biopsies to elucidate clonal evolution. This schematic illustrates the development of a single tumour clone that diverges into multiple distinct entities as a result of selective pressures from cancer progression and drug therapies. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer [1], 2017.
1.2.1 Earlier diagnosis of disease

Tumour material obtained from the blood can be used as a marker of disease for early diagnosis. Dawson et al. have shown that mutations in ctDNA have greater correlation with changes in tumour burden than other blood-based markers, including protein antigens or CTCs [26]. Recently, evaluation of ctDNA in 200 patients with early-stage cancers found mutations in the plasma of 71% of colorectal, 68% of ovarian, 59% of lung, and 59% of breast cancer patients with a minimum threshold of 0.1% allele frequency [27]. Interestingly, presence of ctDNA can also be detected in healthy individuals [29, 27], some of whom progress to develop cancer at a later time [30]. While mutations were detectable in plasma on average 18 months before diagnosis, caution should be used when screening healthy individuals [30, 31]. Mutations in genes commonly linked to cancer, such as TP53, have been identified at low frequencies in the plasma of healthy individuals [32].

As cancer arises from accumulation of alterations, pre-cancerous mutations may be present years before onset of clinical symptoms [33]. Recurrent leukemic mutations accumulate in peripheral blood cells of healthy individuals with increased age. Somatic mutations in genes associated with leukemia have been observed in 1% of people under 50 years of age. Although the prevalence of these alterations increased to 10% in people over 65 years of age, the absolute risk conversion into hematologic cancer was 1% per year as not all mutations led to the development of disease [34]. To distinguish individuals at high risk of developing leukemia, pre-leukemic mutations could be evaluated in peripheral blood of individuals prior to diagnosis of disease. As early mutations are likely low in frequency, a sensitive detection method is required to screen these rare variants.

Since not all mutations in white blood cells (WBC) lead to cancer progression, it is important to distinguish tumour-derived alterations from mutated cfDNA of non-malignant cells. In a study of 821 healthy individuals, there was a high correlation between somatic variants detected between WBC and cfDNA. This warrants caution when searching for low frequency tumour fragments in plasma, as 90% of somatic mutations detected in cfDNA have been identified to be derived from blood cells [35]. Therefore, it is important to use patient matched WBC to differentiate ctDNA from background somatic mutations. To prevent over diagnosis, ctDNA screening could be performed in symptomatic individuals or those at high risk due to germ line genetic carrier status and family history. Early detection prior to metastatic spread can improve patient outcomes, as localized cancer is more amenable to curative therapy.
1.2.2 Minimal residual disease and recurrence monitoring

Following surgery or treatments with curative intent, there may be minimal residual disease (MRD) that is radiologically undetectable. Currently, disease burden is monitored using blood-based markers that lack specificity or imaging which may expose patients to radiation. The paradigm of using ctDNA to detect MRD remains understudied but has shown early promise in certain clinical scenarios for outperforming standard monitoring methods [26, 36]. In a landmark paper, Diehl et al. showed ctDNA as a highly specific marker of recurrence with 100% sensitivity compared to 56% with the standard protein biomarker for tracking disease in colorectal cancer, carcinoembryonic antigen. In patients with detectable ctDNA post-surgery, they observed relapse within one year [37]. In a separate cohort of 230 resected stage II colorectal cancer patients, all of the patients with presence of ctDNA relapsed. On the contrary, 91% of patients with absence of ctDNA had recurrence-free survival [36]. Similarly in a study of 55 patients with early-stage breast cancer receiving neo-adjuvant chemotherapy, serial sampling improved recurrence predictions estimating relapse by a median of 7.9 months prior to clinical observation [28]. Recently in a cohort of 40 patients across stages I-III of lung cancer, ctDNA was detectable in the first post-treatment blood sample of 94% of patients with recurrence. ctDNA identification surpassed radiological monitoring in 72% of patients by a median of 5 months [38]. Together, these studies suggest ctDNA is a reliable marker of recurrence with prognostic predictive value months in advance of imaging.

Liquid biopsies could offer an alternative to invasive procedures for MRD detection such as bone marrow aspirates. While the majority of patients with multiple myeloma achieve complete response to treatment, MRD is prognostic of relapse. Currently, the standard method for MRD testing is flow cytometry where malignant plasma cells are enriched from bone marrow samples and detected using fluorescent markers. MRD by flow cytometry is determined by the detection of 1 myeloma cell amongst 10^6 bone marrow cells; however, flow sorting of a single biopsy is subject to sampling biases [39]. Recently, our group has demonstrated the utility of a hybrid-capture based liquid biopsy sequencing method for complementary molecular profiling of bone marrow in multiple myeloma. We detected 96% of somatic mutations in cfDNA identified from matched bone-marrow samples with 98% specificity. While we found ctDNA at allele frequencies as low as 0.25%, current ctDNA sequencing methods are inadequate for sensitive MRD detection due to the high prevalence of sequence errors [40]. New molecular methods are needed to suppress artifacts for real-time patient monitoring with serial liquid biopsies. Moreover, earlier detection of MRD can be used as patient selection criteria for adjuvant therapy and raises the possibility of reducing over-treatment in disease free individuals.
1.2.3 Evaluating patient response to treatment

As it is not feasible nor desirable to perform serial tissue biopsies, non-invasive liquid biopsies provide a good alternative to characterize patient response to treatment. A considerable amount of effort has been devoted to identifying mutations in the plasma of difficult-to-biopsy cancers. In non-small cell lung cancer (NSCLC), there is a reduction in allele frequency observed in 92% of ctDNA mutations within two days after surgery [41]. Serial plasma biopsies post-surgery have enabled phylogenetic characterization of relapse in early-stage lung cancer patients following adjuvant chemotherapy [11]. As surgery is not accessible to all patients, epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKI) are used as first-line of treatment. Following the first cycle of targeted therapy, resistance occurs in the majority of patients through secondary mutations in the EGFR kinase domain. Plasma assays offer as an alternative for patients who warrant tissue re-biopsies for confirmation of therapeutic resistance drivers, such as the EGFR T790M mutation. While third-generation EGFR inhibitors have been developed to target these resistance mutations [42], ctDNA analyses have revealed tumour heterogeneity leads to further drug tolerance through selection for T790M wild-type cells [43].

Beyond lung cancer, ctDNA profiling has revealed clonal evolution and resistance to EGFR blockade in colorectal cancer. KRAS mutants emerged with introduction of anti-EGFR antibodies and subdued with their withdrawal. Fluctuating mutation frequencies in ctDNA from multiple challenges of EGFR blockade resulted in partial response, providing molecular evidence for rechallenge therapies [10]. Similarly, clonal evolution in the plasma samples of a patient with metastatic breast cancer reflected the clonal evolution inferred from multi-regional tumour tissues [44]. These studies have demonstrated that clonal mutations occurring early in tumour development are ideal candidates for monitoring tumour burden in ctDNA, whereas subclonal mutations arising later in phylongey are better to inform treatment decisions. Currently, the US Food and Drug Administration and the European Medicines Agency has approved ctDNA analysis as a companion diagnostic to aid with NSCLC patient selection for EGFR inhibitors [45, 46]. While these approved tests enable sensitive detection of a few mutations, a challenge still remains for simultaneous profiling of a large number of genomic regions from a limited sample of blood.
1.3 Challenges with low-allele-frequency variant detection

Detection of low-allele-frequency variants has wide implications in oncology from early detection of disease to monitoring advanced-stage malignancies. These rare variant alleles are often confounded by tumour heterogeneity and contamination from normal cells [23]. While this is a known issue for identification of subclones and elucidation of clonal evolution, this also has impact on circulating tumour DNA (ctDNA) detection. In this study, we have developed a method to improve low-allele-frequency variant detection with the main intention of improving ctDNA analysis for liquid biopsies. From sample collection to processing, pre-analytical barriers may diffuse ctDNA signals with noise from germ line cfDNA. The detection limit of low abundance populations is dictated by amount of input molecules, sequencing depth, and level of technical artefacts [47]. Errors can arise as a result of oxidative damage, polymerase mistakes, or sequencer artifacts during DNA extraction and target capture (Figure 1.3) [48, 49, 50]. Collectively, these errors determine the threshold to which we can discern true genetic variants from false positives.

1.3.1 Pre-analytical barriers

Abundance of tumour fragments in the plasma is influenced by disease burden and tumour location. As such, cancers with low mutational burden or those stemming from protected regions like the brain have limited quantities of ctDNA identified in the peripheral blood [24]. One of the major factors confounding ctDNA analysis is contamination of germ line DNA from white blood cell lysis after blood draw. This can influence the limit of detection as mutant allele frequencies can be below one mutant molecule per 10,000 wild-type molecules in advanced-stage solid malignancies (0.01%) [37, 32, 29]. Standardization of sample processing is important for reproducibility and sensitive detection of low-abundance ctDNA molecules. The state of blood following collection impacts the quantity of cell-free DNA as coagulated serum samples can be subject to contamination from lysed immune cells during the clotting process. To improve ctDNA signal, samples should be stored in tubes coated with an anticoagulant, as plasma samples that are free of clotting factors contain lower levels of germ line contamination. It is also important to use an anticoagulant agent compatible with polymerase chain reaction (PCR), such as ethylenediaminetetraacetic acid (EDTA), as other agents have been reported to interfere with PCR efficiency [51].
To prevent contamination by cfDNA, plasma should be isolated through centrifugation within a few hours of blood draw. cfDNA inflation has been observed after 2 hours of storage [51]. This can be problematic in clinical practice as collection sites are often not equipped for immediate plasma isolation. In addition, sample storage conditions and laboratory hours of operation may delay processing and compromise sample integrity. One solution is to use collection tubes containing fixative agents to preserve cell stability and prevent lysis, such as propriety fixatives employed by commercialized Streck tubes. This has been shown to maintain cfDNA yield and conserve background mutational rates for up to 5 days at room temperature [52]. After centrifugation, a layer of buffy coat can be isolated and used as a source of matched germline DNA to determine somatic mutations in cfDNA. The buffy coat is also the source for detecting blood-based cancers present in the circulation, such as acute myeloid leukemia. In addition, there is a multitude of extraction methods available that can impact cfDNA including affinity column, filtration, magnetic bead, phenol-chloroform, and polymer based methods [53, 1]. As there may be size differences between cfDNA and ctDNA, variability between fragment size captured by each method may affect molecular recovery of ctDNA [21].

1.3.2 Technology platforms for rare variant analysis

Analysis of low frequency alleles can pertain to quantification of a single mutation to evaluation of a whole genome. This is relevant for characterizing ctDNA, peripheral blood mononuclear cells, and tissues. Using allele-specific primers for recurrent hot-spot mutations, quantitative PCR (qPCR) can be used to detect ctDNA in the plasma of cancer patients. While some qPCR assays have been approved for clinical use to determine treatment eligibility in NSCLC patients, this method has poor differentiation of low-allele-frequency mutations with 0.05% limit of detection [1]. Subsequent development of digital PCR (dPCR) has allowed for more accurate identification and absolute quantification of rare mutant fragments. As molecules are partitioned into hundreds or millions of parallel PCR reactions and quantified with sequence-specific fluorescent probes, dPCR negates the impact of amplification inefficiencies inherent with conventional PCR [54]. While dPCR can achieve limits of detection below 0.001%, it is difficult to multiplex or simultaneously quantify multiple sequences of interest as each target requires its own unique primer. The possibility of primers hybridizing with one another restricts the applicability of dPCR to at most 10 mutations for simultaneous interrogation [1, 55].

With the development of NGS technologies, a wider range of the genome can be examined at a higher throughput. While whole genome sequencing profiles the entire genome,
targeted sequencing uses panels containing genes or select gene regions for characteriza-
tion of a few exons to the entire exome. Smaller panels targeting fewer number of loci
allows for deeper sequencing at a lower cost. One configuration is amplicon-based assays
where targets are enriched through PCR amplification followed by deep sequencing for
detection of mutations across multiple loci [56]. Another approach is hybrid capture
where sequences of interest are hybridized to biotinylated probes and isolated using mag-
netic beads enabling interrogation of hundreds of genes [32]. Although amplicon-based
protocols require less input DNA and have higher on-target levels, they are prone to
amplification biases, high duplicate rates, and are unable to detect fusions without prior
knowledge of the genomic breakpoint. Hybrid capture on the other hand retains addi-
tional sequences surrounding regions of interest giving potential for discovery of novel
translocations and other structural variants [57].

Whole genome sequencing of cfDNA has identified cancer-specific chromosomal aber-
rations and rearrangements [58]. Recently, shallow whole genome sequencing (sWGS)
of cfDNA with 0.4x mean coverage found aberrations previously unseen in the tumour
biopsy [59]. This highlights the inadequacy of tissue biopsies and the capacity of cfDNA
to reflect spatial tumour heterogeneity. sWGS may offer a cost-effective alternative to
tissue copy number calling. Whole exome sequencing of cfDNA revealed high concor-
dance with tumour biopsies, however analysis was limited to patients with >10% cfDNA
tumour content [60]. Detection of low-allele-frequency mutations requires more sensitive
methods such as digital droplet PCR or targeted sequencing. Current commercial gene
panels have a mutation detection limit above 1% as they are restricted by background
effects inherent to the technology [61, 62]. As recurrent mutations are not found in all
patients, personalized cancer profiling with patient-specific panels in combination with
targeted sequencing can further enhance rare variant detection [32]. The threshold to
which mutations can be confidently called is restricted by background sequencer errors.
Recently, advancements in error suppression strategies have led to improved detection of
allele frequencies below 0.01% for targeted sequencing [29].

1.3.3 Library complexity and artefactual errors

Complexity of a sequence library is reflective of the molecular diversity within a sample
and mirrors the technical challenges throughout the preparation process. Diversity is
characterized by the unique molecules present in the source material. In contrast, molecu-
lar recovery is the proportion of those original molecules retained through the process
of library construction, sequencing, and bioinformatic analysis (Figure 1.3a). During
sample isolation, contamination from exogenous DNA or germ line DNA can mask low-allele-frequency oncogenic variants. This is problematic for analysis of rare fragments such as subclones in mixed cell populations and ctDNA in peripheral blood plasma.

During library preparation, adapter ligation is critical for multiplexing a large number of samples on a single sequence run. The adapter-to-DNA ratio is crucial as insufficient adapters will lead to ligation inefficiencies and excessive adapters could result in adapter dimers, both of which would reduce complexity [63]. While PCR is important for amplification, uneven coverage of molecules as result of GC biased regions reduces the fidelity of the original diversity [64]. Molecular complexity is further impacted by hybridization capture as only select molecules within the target range are retained. Probes for target enrichment of genes are designed based on a reference sequence, thus sequences with substantial deviations or insertions and deletions may not be captured as efficiently [63]. Together, these technical challenges impact molecular recovery, and a method for accurate quantification is needed to characterize library complexity of limited input samples, such as cfDNA.

Sequence errors and technical artefacts pose a major barrier for sensitive detection as they mimic low-allele-frequency somatic variants. Since the limit of detection is determined by the level of background errors, it is important to consider sources of bias during experimental design (Figure 1.3a). Errors in DNA sequence can arise from sample degradation during formalin fixed preservation of tissue. In addition, oxidative damage during DNA extraction or target capture is a prevalent source of sequence errors confounding rare variant detection. 8-oxoguanine (8-oxoG) lesions during DNA shearing or hybrid-capture enrichment results in characteristic imbalance between C>A / G>T substitution errors. Through PCR amplification, oxidation induced 8-oxoG is corrected to a thymine and paired with an adenine, resulting in an asymmetric artefact on one strand of a molecule [65, 29, 49].

Errors can arise during PCR and sequencing. High fidelity DNA polymerases cause 1 error per 3.6 x 10^6 nucleotide incorporations [66]. Error rates inherent to distinct NGS platforms range from 0.1-1% [67]. Sequencer artefacts can arise during cluster amplification, cycle sequencing, or image analysis. While sequencing is a known source of artefactual error, its impact on molecular complexity is less clearly defined. Altogether, these background errors set the threshold to which we can discern true somatic variants and effective error suppression strategies can lower our limit of detection.
Figure 1.3: Stages impacting molecular complexity and artefactual errors. a) Detection of ctDNA is impacted by molecular complexity and is limited by level of background errors. While adapter ligation and target capture may impact molecular complexity, the lower limit of detection is determined by the rate of artefactual errors. These sequence mistakes may arise from polymerase errors, sequencing blunders, or oxidative damage during sample extraction and target capture. b) Raw sequence reads are not representative of true molecular diversity due to PCR duplicates. De-duplication using genome coordinates is not an accurate depiction either as independent molecules may map to the same position. UMIs enable more accurate quantification of molecular complexity and can enable sequence error suppression by correcting artefacts through consensus formation. Although traditional barcoding methods rely on duplicate reads, we have developed a strategy to enable error suppression in single reads.
1.3.4 Molecular quantification and error suppression strategies

Molecular complexity can be measured by the frequency of unique molecules present in sequencing data. This requires removal of redundant sequencing reads from the same template molecule through bioinformatic processing. One approach is to eliminate duplicate copies sharing the same mapping coordinates to a reference genome. However, this is an inaccurate measure as the chance of profiling independent molecules sharing the same genome coordinate increases with sequencing depth (Figure 1.3b). Quantification of molecular complexity using NGS can be improved with unique molecular identifiers (UMIs). Each unique DNA molecule is labeled with a barcode prior to amplification. As each barcode marks a different molecule, reads sharing the same UMI are inferred to result from PCR duplication from the same original template molecule. Thus, quantification of distinct UMIs can determine the absolute number of molecules [68]. In addition, UMIs can be used to aggregate duplicate reads from the same fragment into a consolidated sequence with the most frequent base represented in each position. With these properties, UMIs have been used in a wide range of applications from tagging unique transcripts in single cell sequencing to error suppression in duplex sequencing [69, 70, 71, 29, 27].

A promising strategy to suppress artefactual errors is duplex sequencing. Barcodes are ligated onto the ends of double-strand DNA fragments to track individual molecules for quantification of true molecular complexity. Reads sharing the same UMI are derived from the same strand of a unique fragment and can be collapsed to form single strand consensus sequences (SSCS). This process removes PCR duplicates by establishing a consolidated consensus sequence representative of all duplicate reads. As the most common base is identified at each position, this process eliminates miscalled bases caused by polymerase and sequencer errors. These condensed sequences can be subsequently combined with their complementary strand into duplex consensus sequences (DCS). This additional layer of error removal corrects for DNA damage on asymmetric strands accrued prior to PCR, for example due to oxidative damage (Figure 1.3b). Only mutations present in the consensus sequences of both strands of DNA are retained through this second step of correction. While the duplex strategy has shown greater error suppression, it is an inefficient process dictated by quantity of input molecules and sequencing depth [71, 29].

Barcode-based strategies were initially employed to improve counting accuracy of unique molecules and later expanded to suppress PCR and sequencing errors on single-strands of DNA [72, 73, 74, 68, 75]. In 2012, Schmitt et al. introduced a duplex strategy using UMIs to track and correct errors within double-stranded DNA [70]. As they used a long 12 base barcode on each adapter end, the applicability of their method was restricted to small genomes due to costs associated with long barcodes. Succeeding their
initial work, they re-designed their adapters and demonstrated broader implementation of
the method with targeted sequencing [71]. Subsequently, Newman et al. highlighted the
use of short 2 base barcodes combined with genome coordinates as cost-effective UMIs for
ctDNA. In addition, they combined duplex sequencing with a background error polishing
strategy. While modeling position-specific errors using control samples enables further
error correction, the power of this method relies on additional sequencing of a group of
healthy individuals. Moreover, the use of short barcodes may suffer from UMI collisions
where two molecules could share the same identifier [29]. Recently, Phallen et al. vali-
dated the utility of short barcodes combined with genomic coordinates for detection of
early stage cancers using ctDNA. In addition, they presented different variant filtering
strategies depending on the number of supporting reads for a given molecule. However,
these UMI aware callers do not address common issues filtered from conventional callers
such as triallelic sites, variants observed in control, or clustered positions to name a few.
As duplicate reads are needed to construct consensus sequences, current molecular bar-
coding strategies rely on a minimum of 2 supporting reads. With inadequate sequencing,
consensus based error suppression cannot be achieved for every molecule resulting in a
high abundance of error prone single reads.

1.4 Rationale and hypothesis

In order to achieve accurate molecular quantification and optimal error suppression for
sensitive detection of low-allele-frequency variants, there is a need to overcome barri-
ers imposed by molecular recovery and technical artefacts from library preparation and
sequencing. Current error suppression strategies using UMIs are limited to a subset of
molecules with duplicate reads. I hypothesize that extension of error suppression to single
reads will improve detection sensitivity of rare mutations.

1.5 Objectives

1. Develop an NGS strategy for accurate measurement of molecular complexity.

2. Improve duplex sequencing through error suppression incorporating single reads.

3. Evaluate ability to detect rare variants with sensitive profiling of cell line dilution
   series.
Chapter 2

Materials and Methods

In order to achieve accurate molecular quantification and optimal error suppression for sensitive low-allele-frequency variant detection, we developed a novel computational pipeline for molecular barcode-based sequencing across two applications. The first is LgMid which employs a large (1.2 MB) panel sequenced to moderate depth (4,223x target coverage). The second is SmDeep which employs a small (0.01 Mb) panel sequenced to great depth (186,312x).

2.1 Target panel design

We constructed hybrid capture panels targeting a range of genomic footprints from 13 KB to 1.2 MB sequenced to an average of 186,312x and 4,223x target coverage, respectively. The deeply sequenced small panel (SmDeep) encompassed exons of 5 genes (KRAS, NRAS, BRAF, EGFR, and PIK3CA) important in the mitogen-activated protein kinase (MAPK) pathway. An application of this panel was to assess whether sequencing of cell-free DNA in blood was comparable to bone marrow biopsies for molecular profiling of multiple myeloma [40]. We also moderately sequenced a large panel (LgMid) comprised of 260 leukemia associated genes. This was used for the identification of pre-leukemic mutations in white blood cells of individuals who later developed acute myeloid leukemia. To evaluate our limit of detection using these target panels of different sizes, we mixed cancer cell lines with known genetic alterations to emulate varying levels of mutant allele frequencies from 100% to $10^{-4}$% (Table 2.1). As background errors set the threshold to which we can discern true mutations from technical artefacts, we developed a sequencing strategy enabling error suppression for single reads.

2.2 Next-generation sequencing library preparation

Illumina NGS libraries were prepared for limit of detection assays using cell line DNA. Library preparation was performed according to sample and application as outlined in Table 2.1. Cell line genomic DNA was used to create a 6-point dilution at ratios of 1/5 for LgMid and an 8-point dilution at ratios of 1/10 for SmDeep. For each dilution,
### Table 2.1: Summary of library preparation specifications for each cell line dilution.

<table>
<thead>
<tr>
<th></th>
<th>LgMid</th>
<th>SmDeep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>LgMid</td>
<td>SmDeep</td>
</tr>
<tr>
<td>Cell line spike-in</td>
<td>MOlM13</td>
<td>HCT116</td>
</tr>
<tr>
<td>Cell line background</td>
<td>SW48</td>
<td>MM15</td>
</tr>
<tr>
<td>Dilution points</td>
<td>6 (x2 technical replicates)</td>
<td>8</td>
</tr>
<tr>
<td>Dilution (%)</td>
<td>5 to 5(^2)</td>
<td>10 to 10(^{-4})</td>
</tr>
<tr>
<td>DNA input (ng)</td>
<td>100</td>
<td>60</td>
</tr>
<tr>
<td>Fragment size (bp)</td>
<td>250</td>
<td>180</td>
</tr>
<tr>
<td>Pre-capture PCR cycles</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Samples pooled per capture</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Target Panel</td>
<td>xGen® AML Cancer Panel v1.0</td>
<td>All exons of BRAF, EGFR, KRAS, NRAS, PIK3CA</td>
</tr>
<tr>
<td>Panel size (MB)</td>
<td>1.20</td>
<td>0.01</td>
</tr>
<tr>
<td>Post-capture PCR cycles</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Sequencer</td>
<td>HiSeq 2500</td>
<td>HiSeq 2000 v3</td>
</tr>
<tr>
<td>Sequence length (bp)</td>
<td>125</td>
<td>100</td>
</tr>
<tr>
<td>Number of reads (billions)</td>
<td>5.32</td>
<td>2.05</td>
</tr>
<tr>
<td>Mean Target Coverage</td>
<td>4,223</td>
<td>186,312</td>
</tr>
<tr>
<td>On/Near target</td>
<td>90%</td>
<td>60%</td>
</tr>
</tbody>
</table>

60 and 100ng DNA was sheared before library construction using a Covaris® M220 sonicator (Covaris, Woburn, MA, USA) for 180 and 250 bp fragments in SmDeep and LgMid respectively. The DNA libraries were constructed using the KAPA Hyper Prep kits (#KK8504, Kapa Biosystems, Wilmington, MA, USA) with custom double strand duplex molecular barcodes (xGen Lockdown Custom Probes Mini Pool, Integrated DNA Technologies, Coralville, IA, USA) designed by Dr. Scott Bratman. Following end repair and A-tailing, adapter ligation was performed overnight using 100-fold molar excess of molecular barcode adapters. Agencourt AMPure XP beads (Beckman-Coulter) were used for library clean up and ligated fragments were amplified between 4-8 cycles using 0.5µM Illumina universal and sample specific index primers.

### 2.3 Target capture and sequencing

For each dilution, half of the indexed Illumina libraries were pooled together in a single capture hybridization (Table 2.1). Following the IDT Hybridization capture protocol, each pool of DNA was combined with Cot-I DNA (Invitrogen) and xGen Universal Blocking Oligo (Integrated DNA Technologies, Coralville, IA, USA) to prevent cross
hybridization and minimize off-target capture. Samples were dried and re-suspended in hybridization buffer and enhancer. Target capture with custom xGen Lockdown Probes (Integrated DNA Technologies, Coralville, IA, USA) was performed overnight. Streptavidin-coated magnetic beads were used to isolate hybridized targets according to manufacturer’s specifications. Following target selection, the captured DNA fragments were enriched with 10-15 cycles of PCR. Pooled libraries were sequenced using 100-125 bp paired-end runs on Illumina platforms (HiSeq v3 2000, HiSeq 2500) at the Princess Margaret Genomics Centre. See Table 2.1 for details.

2.4 Data preprocessing

After sequencing, reads were de-multiplexed using sample specific indices into separate paired-end FASTQ files. A two base pair molecular barcode and a one base pair invariant spacer sequence were removed from each read. A thymine base was encoded in the third position for adapter ligation and a spacer filter was enforced to remove reads in compliant with this design. The extracted barcodes from paired-end reads were grouped and written into the FASTQ sequence identifier header of each read for downstream in silico molecular identification [71]. FASTQ files were mapped to the human reference genome hg19 using BWA (v 0.7.12) [76], processed using the Genome Analysis ToolKit (GATK) IndelRealigner (v 3.4-46) [77], and sorted by genome position and indexed using SAMtools (v 1.3) [78]. This process created sorted BAM files containing sequence alignment data.

2.5 Molecular barcode analysis

2.5.1 Unique molecular identifiers

A range of barcode lengths have been developed for the identification of unique molecules. Kennedy et al. originally proposed the concept of duplex sequencing using a 12 bp identifier tag with a 5 bp invariant spacer on each end of a molecule [71]. Subsequently, Newman et al. demonstrated the utility of dual end 2 bp barcodes and 2 bp spacers partnered with genome mapping positions [29]. While shorter barcodes are more cost effective, they are subject to barcode collisions where the same identifier may be assigned to two different molecules. As there are only 256 possible combinations at a given locus with 4 bp barcodes (2 bp on each end), Newman et al. estimated the fraction of molecular loss to be between 0.15% and 0.56% [29]. Phallen et al. further confirmed the utility
of short sequence tags through Monte Carlo simulations of varying barcode lengths in combination with genomic positions. Their evaluations revealed that barcodes between 4-16 bp in length are sufficient to distinguish individual molecules when paired with genome positional identifiers [27].

**Figure 2.1:** UMIs: A barcode family is defined by barcodes and sequence features that enable unique identification of an individual molecule. Our 1) 4-bp barcode allows for 256 different molecules at each 2) genomic position. In addition, 3) CIGAR strings can provide additional identifier diversity as they encode sequence variation to the reference genome highlighting distinct attributes such as clipped regions, insertions, or deletions. Furthermore, read 4) orientation and 5) number can differentiate palindromic barcodes and provide proper molecular grouping of reads belonging to different strands. R1 = Read 1 and R2 = Read 2 of paired-end reads.

Short oligonucleotide barcodes have the benefit of reduced cost for barcode synthesis and conservation of DNA bases for biological DNA in short read sequencing. To characterize unique molecules, we utilized a pair of 2 bp barcodes (4 bp in total) in combination with 4 sequence features from paired-end reads: i) genomic position, ii) concise idiosyncratic gapped alignment report (CIGAR), iii) read orientation, and iv) first read or second read (Fig. 2.1). Hybridization capture approaches have the benefit of catching a wide range of molecules with varying mapping positions, whereas amplicon-based methods capture fragments with conserved positions. By utilizing the diverse genome locations of hybrid capture fragments, shorter barcodes can be employed in combination for unique molecular identification. By encoding both the i) start and end locations of each fragment, we can preserve read pairs across large insert sizes of chromosomal translocations. A 4-bp barcode enables $4^4$ or 256 unique possible combinations with each positional pair. In addition, we integrated ii) CIGAR strings in our identifier as they reflect alignment to the reference genome (Fig. 2.2). CIGAR operations can add diversity through incorporation of read features such as M = match/mismatches, I = insertions, D = deletions, S = soft clips, and H = hard clips. Inclusion of sequence alignment properties may improve identifier complexity and expand the pool of UMI combinations. This can potentially
enable more accurate categorization of reads and mitigate barcode related errors.

**Figure 2.2:** Source of CIGAR: When DNA sequencing reads (grey bars) are aligned to a reference sequence, specific genome coordinates and mismatches to the reference (CIGAR strings) are noted in the sequence alignment file. CIGAR strings encode information about fragment clipping (S, soft clipped bases in 5’ or 3’ end of the read that are not part of the alignment), as well as sequence concordance with the reference genome (M, match/mismatch bases), insertions (I), and deletions (D).

Moreover, we incorporated iii) strand orientation and iv) first or second sequence in paired reads to distinguish palindromic barcodes for differentiation of reads from separate strands. As there are \(4^{n/2}\) (\(n = \) number of bp in barcode) palindromic combinations, our 4 bp barcode (e.g. AT from read R\(_1\) and AT from R\(_2\)) encodes 16 possibilities where the barcode sequence reads the same for plus and minus strands. Without these sequence distinctions, reads from separate strands of a molecule may be grouped together. While this only represents 6.25% of our barcodes, this is an issue that increases exponentially with longer barcodes where genome coordinates alone are insufficient for differentiating palindromic sequences. Thus, it is crucial to employ multiple sequence measures to supplement short barcodes for accurate duplex molecular identification.

### 2.5.2 Single strand consensus sequences

Using our UMIs, reads derived from the same strand of a molecule can be condensed into single strand consensus sequences (SSCS) (Fig. 2.3). First, a filter was applied to exclude reads which were unmapped, paired with an unmapped mate, or had multiple alignments. Paired reads were assigned UMIs as described above using barcode, genome
Figure 2.3: Duplex sequencing schematic: An uncollapsed BAM file is first processed through SSCS_maker.py to create an error suppressed single strand consensus sequences (SSCS) BAM file and an uncorrected Singleton BAM file. The single reads can be recovered through singleton_strand_rescue.py, which salvages singletons with its complementary SSCS or singleton. SSCS reads can be directly made into duplex consensus sequences (DCS) or merged with rescued singletons to create an expanded pool of DCS reads (Figure illustrates singleton rescue merged work flow).
mapping, CIGAR string, strand of origin, orientation, and read number information. Reads sharing the same UMIs were grouped into the same read family. Only families with 2 or more members were error suppressed and collapsed to form SSCSs as following:

1. For each position $i$ across a sequence length, a Phred quality threshold of Q30 was enforced for every read $r_j$. As Phred scores are a measure of the quality of base assignment called by the sequencer, only bases with an error probability of one in a thousand or less ($r_{i,j} >$Q30) were evaluated for consensus formation.

2. The most frequent base at each position across all replicate reads of the same molecule was established as the consensus ($c_i$). The most common base was assigned if the proportion of reads representing that base was greater than or equal to the threshold required to confidently call a consensus (default cutoff 0.7 - based on previous literature[71]), otherwise an N was assigned.

3. Here, we introduced a molecular Phred quality score for each position to reflect bases that went into making each consensus sequence. A sequencer Phred score ($Q_{seq}$) is a measure of the estimated probability ($P_{seq}$) of incorrectly calling a single base:

$$P_{seq} = 10^{-Q_{seq}/10}$$

$$Q_{seq} = -10\log_{10}(P_{seq})$$ (2.2)

A molecular Phred score ($P_{mol}$) estimates the probability ($P_{mol}$) of incorrectly calling $c_i$. It is the product of error probabilities of Phred scores ($Q_k$) corresponding to bases supporting $c_i$, where $k$ is the index of scores and $n$ is the total number of scores matching the consensus base.

$$P_{mol} = \prod_{k=1}^{n} 10^{-Q_k/10}$$ (2.3)

$$Q_{mol} = -10\log_{10}(P_{mol})$$ (2.4)

The molecular Phred score can be simplified to the addition of quality scores to summarize probability of errors.

$$Q_{mol} = \sum_{k=1}^{n} Q_k$$ (2.5)

While Phred quality scores can theoretically range from 0 to infinity, a maximum Q60 cap was enforced to accommodate thresholds imposed by various genomic tools
to flag misencoded quality scores. If no consensus could be reached at a position and an N was present, $P_{mol} = 0$ was assigned for that base.

4. As each SSCS represents multiple reads derived from the same strand of a unique fragment, a consensus query name was assigned to each SSCS pair. Similar to our UMIs, the pairing tag consists of a barcode along with 4 sequence features: i) genome mapping ordered by coordinate, ii) strand of origin inferred from read orientation and number, iii) CIGAR string ordered by strand of origin and read number, and iv) family size (number of reads supporting SSCS).

Filtered reads, SSCS with two or more supporting reads, and singletons were written as separate BAM files. While single reads are traditionally excluded from consensus formation, we incorporated a singleton rescue strategy enabling error suppression without molecular duplicates. Salvaged singletons can be combined with SSCS for downstream DCS formation.

### 2.5.3 Singleton rescue (SR)

Previous UMI-based techniques for duplex sequencing were restricted by the high depth required for duplicate recovery. Kennedy, Newman, and Phallen consensus approaches have limited capacity for error correction in singleton dominant samples [71, 29, 27]. Here, we introduce two approaches for singleton rescue using the duplex nature of DNA molecules for biological elimination of technical artefacts (Fig. 3.2). Following the formation of SSCS, singletons can be corrected with its reverse complement SSCS for i) singleton rescue by SSCS. If a complementary SSCS cannot be identified, single reads can also be paired with its reverse complement singleton for ii) singleton rescue by singletons. In this case similar to the duplex consensus strategy, only two reads corresponding to the dual strands of a template molecule are required to error correct singletons as following:

1. UMIs were assigned to singleton and SSCS reads. For each singleton, a duplex identifier was determined by interchanging barcodes, swapping CIGAR strings, switching strand direction, and changing read number. If $R_1$ and $R_2$ on a positive strand had AC/GT as barcodes and 98M/4S94M for CIGAR strings, their duplex would be GT/AC for barcodes and 4S94M/98M for CIGAR strings on the minus strand. $R_1$ in the forward orientation on the plus strand corresponds to $R_2$ in the forward orientation on the minus strand.
2. Singleton rescue can be achieved using either its reverse complement i) SSCS or ii) singleton read. For each base, a Phred quality filter of Q30 was enforced to remove error prone bases. Consensus sequences were formed by taking concordant bases at each position or by assigning Ns for mismatches. Similar to SSCS, molecular Phred quality scores were computed by taking the sum of sequencer Phred qualities and capped at a maximum score of Q60.

3. Error suppressed singleton pairs were assigned a consensus query name as described above for SSCS reads. Recovered singletons were written to separate BAM files depending on method of rescue. They were subsequently merged with SSCS reads for duplex formation. The inclusion of singletons allows a larger pool of reads to undergo duplex error correction removing strand bias errors.

### 2.5.4 Duplex Consensus Sequences

For optimal error suppression, duplex consensus sequences (DCS) can be established by comparing SSCSs corresponding to the double strands of a molecule. This second stage of duplex error suppression eliminates asymmetric strand damage induced by oxidation. While DCS methods form sequences with lower rates of error, it has been characterized as a fairly inefficient process \[71, 29\]. Here, we optimized DCS formation by expanding our pool of candidates to include singletons. While our method error corrected singletons prior to DCS formation, uncorrected singletons can be used directly for DCS in theory as both DCS and singleton rescue utilize the duplex strategy for error suppression. Consensus sequences were established by preserving matched bases between reverse complementary reads. Mismatches were denoted as N and molecular Phred qualities were assigned to consistent bases. Although DCS reads have the lowest rates of error, they only depict a portion of the total molecular population. To portray accurate molecular representation for variant calling, a BAM file containing all unique molecules can be created by combining DCS, SSCS (without duplex pair), and singletons.

### 2.6 Evaluation of consensus sequence formation

Efficiency of consensus formation reflects the frequency of consensus sequences generated per read. This is determined by the average number of reads required to construct a consensus. For example, an efficiency rate of 10% indicates each read contributes to 0.1 of a consensus or rather 10 reads are needed to form a single consensus.
In order to compare target panels of varying sizes, efficiency rates were calculated using the mean target coverage \( (cov) \). GATK (v 3.6) DepthOfCoverage was used to determine mean fragment coverage per target position. Notably, we performed fragment counting as it considers overlapping reads as a single entity rather than double-counting those reads:

\[
Efficiency = \frac{cov_{DCS}/SSCS}{cov_{Uncollapsed}}
\]  

(2.6)

As DCS formation is dependent on the number of SSCS and rescued singletons, DCS recovery rates were estimated by comparing observed over expected rates:

\[
Recovery_{DCS} = \frac{Observed_{DCS}}{Expected_{DCS}} = \frac{cov_{DCS}}{\left(\frac{cov_{SSCS}}{2}\right)}
\]  

(2.7)

### 2.7 Error analysis

Error is an estimation of the confidence that a single base is assigned to the correct nucleotide. Probabilities of error were determined using the integrated digital error suppression (iDES) tool ([https://cappseq.stanford.edu/ides/download.php#bgReport](https://cappseq.stanford.edu/ides/download.php#bgReport)) [29]. BAM files were first converted to base frequency files for each genomic position using `ides-bam2freq.pl`. With the `ides-bgreport.pl`, background errors were calculated using non-reference bases with at least one read support below 5% allele frequency. Error rates were determined as the number of non-reference bases over all sequenced bases within our capture panel range.

### 2.8 *in silico* downsampling

To compare hybrid capture panels of different sizes sequenced to various depths, we performed down-sampling of average read coverage per target position to set intervals. 9 intervals were established between 128,000x to 500x coverage with depth decreasing by half each step. Each sample was reduced in coverage to the set intervals through *in silico* down-sampling of paired-reads with Samtools (v 1.3). This process was repeated ten times for each sample across all intervals to address sampling biases. Next, each down-sampled file was run through our duplex sequencing pipeline to generate singletons, SSCS, DCS, singleton rescue by SSCS, and singleton rescue by singletons files. Rescued singletons were merged with SSCS to generate SSCS + SR for down stream duplex
formation (DCS + SR). Efficiency for consensus formation and molecular recovery was assessed for each error suppression strategy across the broad range of coverage intervals.

2.9 Mutation analysis

We selected variants based on annotated SNPs from the Cancer Cell Line Encyclopedia (CCLE) overlapping our target panel and corresponding to the cell lines we used for our dilution series. These variants were filtered for only those found exclusively in the spike-in cell line. For SmDeep, we identified 3 mutations in \textit{BRAF}, \textit{KRAS}, and \textit{PIK3CA} of cell line HCT116. Whereas, 5 mutations in \textit{DST}, \textit{NF1}, \textit{RYR1}, \textit{SMG1}, and \textit{VCAN} were called for LgMid.

Somatic mutation calls were generated for each sample using MuTect (v 1.1.5) with the following parameters [79]:

- `--enable_extended_output` -tumor_f_pretest 0.000001f  -downsampling_type NONE -force_output -force_alleles -gap_events_threshold 1000 -fraction_contamination 0.00f -coverage_file

We force called every base of each selected CCLE annotated position to evaluate the number of reads and allele frequency supporting each variant. MuTect metrics were used to assess limit of detection and background noise for each stage of barcode-mediated error correction across the dilution series.
Chapter 3

Results

3.1 Sequence properties can enhance barcode depiction of molecular diversity

Here, we developed a method with a 2 base pairs (bp) barcode followed by a 1 bp spacer for adapter ligation. We designed compact molecular labels to reduce barcode synthesis expenditures, while minimizing artificial sequence uptake reserving more bases for biological DNA. To supplement our short barcodes, we incorporated sequence features including genome coordinates which have been shown to be effective for distinct molecular characterization of ctDNA [29, 27]. In addition, we integrated sequence properties (CIGAR strings) to our UMIs indicative of base concordance, fragment clipping, insertions, and deletions (Fig. 2.1). We assessed the contribution of these additional features for capturing molecular diversity by quantifying reads with misalignment to the reference genome. To evaluate our method, we compared molecular complexity and error suppression in cell line dilutions across two contrasting scenarios: targeted deep sequencing and broad moderate sequencing (SmDeep and LgMid, see methods). As we require two or more reads supporting the same UMI to construct single-strand consensus sequences (SSCS), single reads without duplicate support for error correction were classified as singletons. We observed a higher ratio of alternative alignments within singletons compared to SSCS (Fig. 3.1a), which suggests CIGAR strings may pose as a stringent filter for consensus formation. Next, we wanted to explore whether inclusion of alignment information enables more accurate depiction of molecular complexity and improves error suppression. To test this idea, we compared consensus sequences derived from molecular identifiers with and without alignment information. While there were minor changes in reads of LgMid, there was on average a 33% increase in singletons of SmDeep with CIGAR labeling (Fig. 3.1b). As molecules were saturated across a narrow target range with deep sequencing, these results suggest barcodes and genome coordinates alone were insufficient for characterizing molecular complexity. This may be also be problematic for cfDNA molecules in plasma as their tight size distribution would result in a narrow range of mapping coordinates. Expanding the combinations of molecular identifiers with
alignment information may improve depiction of cfDNA populations. In order to understand the impact of alignment on consensus-based error correction, we evaluated the change in background error profiles by examining non-reference bases excluding germline
SNPs. With CIGAR derived consensus sequences, we observed an average decrease in error rate by 7% in LgMid and 41% in SmDeep. Notably, we found the reduction in SSCS error was correlated with increased singletons for SmDeep (n = 8, r = -0.87, p < 0.005, Pearson correlation). This is likely due to CIGAR identifiers enabling more defined read categorization. As large read families sharing the same barcode and genome coordinate are differentiated into smaller groups with varying alignment, some UMIs may lose multi-read support and become singletons. Overall, our findings suggest alignment information enables more distinct molecular identification and improves error suppression.

3.2 Singleton rescue improves efficiency of consensus formation

A major barrier to adequate error suppression with conventional UMI-based strategies is the constraint of high depth required for sufficient molecular recovery. Traditional duplex approaches rely on duplicate reads to form consensus sequences and have limited utility when there is inadequate coverage and an overabundance of singletons [71]. Here, we propose harnessing the double-strand nature of DNA molecules for biological elimination of technical artefacts in singletons. We can recover singletons by employing the duplex strategy of error correction using a complementary SSCS or singleton (Fig. 3.2). In the scenario where one strand $S_1$ of a fragment has multiple copies and the other strand $S_2$ only has one copy, reads corresponding to $S_1$ are first formed into a SSCS. The consensus sequence of $S_1$ can be subsequently matched with the singleton of $S_2$ for 1) singleton rescue by SSCS. Alternatively if both $S_1$ and $S_2$ are singletons, the lone reads can error correct one another through 2) singleton rescue by singletons. Together, these dual strategies of singleton rescue (SR) increase the candidate pool of reads for downstream duplex formation.

To evaluate the SR strategies, we examined consensus formation in two mixtures of cancer cell lines emulating varying levels of ctDNA. As a consensus sequence requires a minimum of two duplicate supports for construction, 67% of all reads in LgMid (4,223x mean target coverage) qualified for consensus formation (Fig. 3.3). This corresponded to a 25% single-strand consensus sequence formation rate (equation 2.6), which is equivalent to an average of 4 reads used to build each SSCS. With the assumption that 2 single-strand molecules form a duplex molecule, we would expect the frequency of DCS to be half of total SSCS if every double-strand molecule was recovered. However, only 14% of the expected DCS (equation 2.7) were observed. This corresponded to a low efficiency
Figure 3.2: Dual strategies of single read error suppression: Singleton rescue by SSCS starts with collapsing duplicate reads to form SSCS, which are subsequently used to correct the reverse complement singleton. Additionally, singletons can be rescued by other singletons through complementary strand error suppression.

rate of \(\sim 2\%\) for DCS formation with an average of 54 reads needed to construct each DCS. As a third of all sequences lacked duplicate read support, our SR strategy recovered a quarter of singletons. SR corrected 10% of singletons using the complementary SSCS strand and 12% using the complementary singleton strand. This suggests a quarter of singletons had the potential to qualify for duplex correction, but would have been missed with conventional UMI strategies due to the lack of duplicate support. Furthermore, one tenth of those singletons were the complementary strand to SSCSs, which explains the low DCS recovery rate we observed. With the traditional duplex model, these SSCSs failed to form DCSs as their other strand lacked sequencing replicates. Expansion of consensus formation to singletons reduced the number of reads needed to build a consensus sequence to 3 for SSCS (33% efficiency rate) and 11 for DCS (9% efficiency). Together, the dual strategies of SR recovered 53% of expected DCS resulting in a 4.6-fold improvement compared to conventional duplex methods.

In contrast to the panel sequenced to moderate depth, 98.7% of reads in the deeply sequenced SmDeep (186,312x mean target coverage) contributed to consensus development. As the majority of molecules had duplicate read support, SSCS formation was inefficient needing 30 reads on average to assemble a single consensus sequence (3% SSCS formation rate). Consequently, DCS efficiency was even lower at 0.5%, which corresponds to an average of 184 reads used to construct each DCS. The high ratio of duplicates
to singletons sequenced account for the low efficiency rates. However, deep sequencing corresponded to a higher DCS recovery rate of 33.6% compared to the 14% previously observed with sequencing to moderate depth. As singletons accounted for only ~1% of all reads in SmDeep, SR had insignificant impact on SSCS and DCS formation as relatively few singletons were recovered. Although singleton correction improves consensus efficiency, it is not equally effective in all scenarios. Therefore, to efficiently maximize molecular recovery, there is a need to characterize the optimal sequencing context for error suppression.

![Figure 3.3: Proportion of reads with duplicate support eligible for consensus-based error suppression. Singleton rescue (SR) enables additional error correction of singletons through rescue strategies using complementary SSCS or singletons. The left shows the average read distribution of a large panel with sequencing to moderate depth (LgMid, n = 12), while the right shows a summary of reads from a small panel with deep sequencing (SmDeep, n = 8).]

3.3 Singleton rescue achieves comparable rates of error suppression as duplex correction

Singletons typically cannot be error corrected with barcoding and contain artefacts which may confound detection of variant allele frequencies <0.1% expected in ctDNA. To assess the error suppression performance of SR, we evaluated errors at each stage of consensus formation. Approximately half of the large panel was error-free, whereas the small panel had errors in nearly every position (Fig. 3.4). The lower rate of error-free positions in
SmDeep compared to LgMid may be explained by the higher probability of mismatched bases occurring within a narrow target range. Although SSCS reduced the frequency of positions with errors to less than 25%, duplex was required for maximal error reduction. DCS had the highest rate of error-free positions with 99.96% for LgMid and 99.82% for SmDeep. Across both panels, SR achieved comparable metrics with mismatches found below 0.1% across the target panel.

Figure 3.4: Error suppression performance at each stage of error correction between sequencing with moderate (LgMid) and deep (SmDeep) coverage. Background error determined by non-reference bases at allele frequencies below 5%. Error bars represent standard error across all samples.

Panel-wide error rates corresponded with error-free positions, where higher frequency of positions without errors concurred with lower error rates. Across the selected genes of both target panels, we found SSCS reduced error rates to below 0.01%. Whereas, removal of asymmetric-strand damage with DCS suppressed errors to an order of magnitude lower below 0.001% (Fig. 3.4). While SR of SmDeep presented the lowest error rates across the error suppression strategies, this was biased by the relatively low proportion
of singletons rescued. With deep sequencing, only 1% of SmDeep reads were singletons and 2% of those singletons were rescued. As SmDeep did not adequately represent correctable singletons, we focused on LgMid to provide a better assessment of errors with SR. Across the 23% of singletons corrected in LgMid, we observed error rates of $1.0 \times 10^{-3}%$ for SR by SSCS and $1.2 \times 10^{-3}%$ for SR by singletons. Most notably, both SR strategies minimized singleton errors by 25-fold down to an average of $1.1 \times 10^{-3}%$ per base. This is noteworthy as it suggests high quality error suppression can be achieved with an individual read, challenging the fundamental notion of requiring multiple supporting reads for conventional consensus formation. Despite the high rate of artefacts in singletons, our correction strategy obtained similar background levels of error as top performing DCS.

### 3.4 Oxidative damage is effectively suppressed by singleton rescue

An analysis of background artefacts in LgMid across base substitution classes (C$>$A, C$>$G, C$>$T, T$>$A, T$>$C, T$>$G; and their reciprocal substitutions) revealed different error profiles with each stage of consensus correction. We observed similar error rates across all 12 substitution classes between uncollapsed reads and singletons without error suppression, suggesting singletons are the result of molecule sampling biases and are not inherently different in error profiles. Consistent with previous reports by Schmitt et al. and Newman et al., we found errors indicative of oxidative damage in uncorrected and SSCS reads as indicated by the imbalance between C$>$A/G$>$T transversions (Fig. 3.5b) [70, 29]. Oxidation induces 8-oxoguanine (8-oxoG) lesions during DNA shearing and hybridization capture. Through PCR amplification, 8-oxoG is corrected to a thymine and paired with an adenine, resulting in asymmetric damage to one strand of a molecule (Fig. 3.5a).

As our hybrid capture panel utilizes probes generated against the positive strand of the human genome reference sequence, only the plus strand of DNA was captured. Consequently, we observed a ratio of 0.6 G $>$T compared to C$>$A errors. Although Schmitt et al. previously reported a higher frequency of G $>$T errors, our findings are consistent with those previously reported by Newman et al. where they demonstrated inverse ratios in substitution errors depending on the strandedness of probes [70, 29]. While the oxidative damage signature was present in uncollapsed and SSCS reads, we did not observe its presence within DCS (Fig. 3.5c-e). Together, these results demonstrate the
importance of DCS error suppression for removal of asymmetric strand damage induced by oxidation.

Next, we explored the ability of SR to address mutational signatures of error by examining base substitutions. We observed elimination of oxidative damage across both SR by SSCS and SR by singletons (Fig. 3.5f-g). Interestingly, this raises the question of whether multiple reads are necessary for error suppression as complementary singletons achieved error rates and substitution profiles similar to DCS. Both DCS and SR rely on the same mechanism of artefact elimination as they are contingent on consolidation of double-strands. While ratio of substitution errors vary between samples in DCS, this difference is potentially due to the few reads available to form DCS. From ~4,000x uncollapsed reads per target position, only 80x reads remain after DCS construction. Contrary, SSCS rescued 142x singletons and complementary singletons recovered 171x reads per target position. As both strategies of SR was similar to DCS across error-free positions, error rates, and substitution profiles, our findings strongly support inclusion of singletons to maximize error suppression.

3.5 Duplex molecular recovery is maximized through singleton rescue

While UMI based error suppression strategies have been deployed for circulating tumour DNA detection, DCS has been described as an inefficient process [29, 27, 71]. This is partly due to the need for deep sequencing, as published UMI algorithms require a minimum of 2 read supports for consensus formation. Since more replicates are needed to construct consensus sequences, less molecules will meet the criteria resulting in lower rates of SSCS and DCS formation. In addition to the bottlenecks on molecular recovery during library preparation and sequencing, bioinformatic analysis may also bias quantification of molecular diversity depending on criteria for defining unique molecules. Here, we explore the impact of an alternative UMI approach with expansion of error suppression to singletons. As SR has the potential to broaden the applicability of UMI correction, we wanted to determine the conditions in which it can improve SSCS and DCS formation. To compare efficiency of consensus formation across panels of different sizes sequenced to various depths, we performed in silico down-sampling to coverage intervals between 500x and 128,000x read per target base. Each sample was downsampled to 9 coverage levels starting at 500x reads per target position and doubling with each increased interval.

Through ten simulations across nine intervals, we assessed the frequency of SR and
Figure 3.5: Duplex strategies of error correction suppresses oxidative damage. 

a) Schematic of oxidative damaging induced 8-oxoguanine (8-oxoG) asymmetric strand lesions in DNA. 8-oxoG arises from oxidation and converts to thymine after PCR amplification. As only plus strands are captured with our target panel comprised of negative probes, there is a bias for C>A errors.

b) Comparison of 12 substitution classes between uncorrected and error suppressed sequences in LgMid (n=12). c-g) Ratio of errors between reciprocal base substitutions. Imbalances between C>A/G>T is indicative of oxidative damage.

its impact on consensus formation across a broad coverage range. As we utilize the double-strand nature of DNA for biological elimination of errors, the success of SR was
Figure 3.6: Singleton rescue impact on consensus formation across a wide coverage range (500x to 128,000x). Results are shown for LgMid and SmDeep cell line dilutions that were down-sampled by half at each interval step. Error bars represents standard error of all the samples across ten simulations. Mean target coverage is log10 transformed to show trends at lower range. 

**a)** Proportion of singletons rescued through complementary SSCS and singleton correction.  
**b)** Efficiency of SSCS formation as determined by number of SSCS / total uncollapsed reads.  
**c)** DCS efficiency as calculated by number of DCS / total uncollapsed reads  
**d)** DCS recovery rate of observed / expected DCS. Expected DCS corresponds to half of all SSCS reads as two single molecules should theoretically form a DCS. Efficiency is an assessment of over-sequencing relative to unique input molecules, whereas recovery is an estimate of molecular retrieval.

dictated by the presence of its complementary strand. With increased sequencing depth, we observed a greater portion of unique molecules and consequently a higher number of rescued singletons (Fig. 3.6a). Starting from an average of 500x uncollapsed reads per target position, we observed a 3% recovery of singletons in SmDeep. As the number of reads per position doubled at each coverage interval, we found the frequency of SR to double as well. This trend continued until reaching a peak SR rate of 20% at 8,000x coverage, where one out of five singletons could be recovered. Beyond this point, we
observed a decrease in SR suggesting an increased prevalence of duplicate reads. This was validated by SSCS efficiency, where we found the distribution of SR corresponded with SSCS formation. As the development of SSCS relies on repeat measures of the same molecule, rate of consensus formation increased until sign of unique molecular saturation observed around 8,000x coverage (Fig. 3.6b). This inflection marked the maximum efficiency rate of SSCS in SmDeep generated from 27% of molecules in the library, representing an average of 4 reads to construct a single-strand molecule. After this point, we began to saturate our unique single-strand molecules with duplicate reads, which led to decreased SSCS efficiency and number of singletons rescued.

Overabundance of singletons pose as a major limitation for conventional UMI methods as there are insufficient reads to form consensus sequences. At 500x reads per target position, inclusion of SR improved SSCS efficiency from 6% to 8% in SmDeep. With increased depth of coverage, we observed a gradual improvement in SSCS efficiency through SR, reaching a peak of 33% at 8,000x mean target coverage. This is a 6% improvement from conventional SSCS formation where only duplicate reads are utilized. However, increases in SSCS efficiency with SR was only observed until 32,00x mean target coverage in SmDeep, at which point we saw no improvements from SR (Fig. 3.6b). Our results suggest SSCS formation only benefit from SR at sequencing depths where molecules lack duplicate reads.

While we also observed increased DCS formation with higher depth of sequencing, the peak DCS rate of 3% was not observed until 16,000x mean target coverage (Fig. 3.6c). Despite deep sequencing of SmDeep, we still found DCS formation to be highly inefficient requiring a minimum of 33 reads at its highest observed performance. With the inclusion of SR, we observed an increase in DCS efficiency and a shift towards more effective DCS formation at lower coverage. This resembled the distribution of SSCS efficiency, as peak DCS efficiency occurred between 4,000x - 8,000x reads per target base. In addition, expansion of DCS formation to include SR resulted in a 2-fold improvement in DCS efficiency up to 6% in SmDeep. While LgMid could only be down-sampled to coverage intervals between 500x and 4,000x as it was moderately sequenced, we observed similar trends in SR and SSCS/DCS efficiency as SmDeep. Notably, LgMid achieved 8.7% DCS efficiency with an average of 4,000x reads per target position. This corresponded to approximately 12 reads used for DCS construction, which is the highest efficiency reported in literature to date.

Next, we assessed the frequency of expected double-strand molecules to estimate the recovery of molecular complexity after library preparation, sequencing, and bioinformatic analysis. DCS recovery was calculated by taking the ratio of observed over expected DCS.
Since two single-strand molecules could theoretically form a DCS, we determined the expected frequency of DCS to be half of the total SSCS population. Using conventional UMI principles requiring duplicates to construct each consensus sequence, we observed a gradual recovery of DCS with increased depth of coverage. At 64,000x reads per target position, we reached DCS saturation with only 34% recovery from both strands of a DNA molecule (Fig. 3.6d).

Expansion of DCS formation with SR, maximized DCS recovery regardless of sequencing depth. Starting from 500x uncollapsed reads per target position, we demonstrated that incorporation of SR improved DCS recovery rate from 0% to 40% in SmDeep and 50% in LgMid. Notably, the frequency of recovered DCS remained in this range despite increased sequencing coverage. This suggests inclusion of consensus sequences from complementary singletons overcomes traditional hurdles of error suppression and can achieve optimal recovery of DCS at any given sequencing depth.

As not all DCSs were retrieved, our results suggest the matching strand was missing for half of all molecules in SmDeep. Further enhancements in library preparation upstream of in silico analyses may improve double-strand molecular recovery. While both dilutions displayed similar trends across all four performance metrics (Fig. 3.6), the variability between them may be accounted by the difference in library preparation and sample input (see methods). Together, our findings demonstrate the ability of SR to recover additional molecules for error suppression, potentially increasing the sensitivity of mutation detection.

### 3.6 Singleton enabled error suppression improves lower limit of mutation detection

Using mixed cancer cell lines with known genetic alterations, we emulated varying levels of mutant allele frequencies. For each dilution series, we evaluated the frequency of SNPs in the spike-in cell line previously identified by the Cancer Cell Line Encyclopedia (CCLE). We called every base of these CCLE annotated SNPs found within each target panel. In SmDeep, we examined the frequency of three known mutations in \textit{BRAF}, \textit{KRAS}, and \textit{PIK3CA} of cell line HCT116 across the dilution series from 100% to $10^{-4}$%. With deep sequencing, we detected the \textit{BRAF} and \textit{KRAS} mutations down to 1% and \textit{PIK3CA} down to 0.1% dilution in the uncollapsed reads without error suppression. Beyond these frequencies, detection was prohibited by the rate of background errors (Fig. 3.7).

To improve upon this, we used our SSCS error correction and observed suppression
of artefacts represented by alternative mutations that are not matching the reference or known allele. This process permitted identification of all three variants at 0.1% dilution with observed allele frequencies as low as 0.01% for the \textit{BRAF} mutation. While DCS error suppression was required for complete removal of artefacts, known mutations were only uncovered at 1% dilution for \textit{BRAF} and \textit{KRAS} and 10% for \textit{PIK3CA}. While the \textit{PIK3CA} mutation was detected at a lower dilution compared to the other two mutations in uncollapsed reads, the absence of the \textit{PIK3CA} mutation in lower dilutions of DCS suggests poor double-strand recovery. These results highlight the variability in molecular recovery and the need to optimize experimental design for improved double-strand molecular retention as currently only 33.6% of maximum theoretical DCSs were retained.

As nearly all of the reads in SmDeep had duplicates sequenced, few molecules were recovered with SR. Consequently, we did not identify observable changes in our limit of detection with the addition of error suppressed singletons (Fig. 3.9a). Without elimination of artefacts, singletons contained errors corresponding to alternative bases rather than the reference or known allele. Despite error correction using duplicate reads, a few errors still persisted in SSCS. As these artefacts in SSCS were only found in 1-3 reads, filters can be imposed to enable higher confidence variant calling with SSCS (Fig. 3.9b).

In the LgMid experiment, we evaluated 5 mutations present in the MOLM13 cell line that were captured by our mutation panel (\textit{DST}, \textit{NF1}, \textit{RYR1}, \textit{SMG1}, \textit{VCAN}). With a dilution from 100% to 0.04%, all 5 mutations were detectable at the 1% spike-in in the uncollapsed reads (Fig. 3.8). Although the mutation in \textit{DST} was also identified at 0.04%, it was undetectable at the 0.2% dilution in uncorrected reads. With SSCS, \textit{DST} and \textit{RYR1} mutations were detectable down to 0.04%. While \textit{NF1} was also identified at similar frequencies, it was also found in the background cell line along with an alternative allele, both of which were artefacts. As artefactual errors compete with signal of true variants in SSCS, it is difficult to confidently call such mutations in \textit{NF1} as errors are retained at 1% allele frequency. Although detection limit of \textit{SMG1} was the same in uncollapsed and SSCS, the mutation in \textit{VCAN} was detected at a lower limit of 0.2% allele frequency in SSCS.

Despite DCSs presenting the lowest error rates, poor DCS recovery with sequencing to moderate depth resulted in reduced number of mutations identified at lower dilutions compared to uncollapsed reads. Incorporation of SR had minimal impact on mutation detection in SSCS, whereas inclusion of SR with DCS improved mutation detection from 5% to 1% in \textit{SMG1} and from 1% to 0.2% in \textit{VCAN}. In addition, expansion of DCS with singletons recovered the known mutation in \textit{NF1} at the 1% dilution, which was previously unidentified in DCS. Together, our results demonstrate the utility of SR to improve DCS
recovery enabling high confidence mutation detection down to 0.2% frequency in DCS.

In LgMid, we observe frequencies of background error at 0.1% in uncollapsed reads prohibiting detection of mutations below such threshold (Fig. 3.10a). While the reduction in errors enabled variant detection at allele frequencies down to 0.04% in SSCS, false positives still persisted as observed by the presence of the spike in variant at 0% dilution. Notably, these false positives were removed in DCS where no sign of the spike-in mutant was observed in the undiluted background cell line (Fig. 3.10b). While DCS of SmDeep was completely error free, a single read supporting an alternative mutation indicative of error was present in DCS of LgMid at the 1% dilution. As a single error persisted into DCS, a lenient filter could also be applied to DCS for higher confidence variant calls. While DCS may enable higher confidence calling, SSCS achieves a lower limit of variant detection. In order to optimally utilize error suppressed consensus sequences, our results suggest a UMI aware caller is needed to enable different filtering strategies for SSCS and DCS depending on confidence of sequence quality. Polishing strategies to model background error rates, as previously described by Newman et al., may further improve our limit of detection, as it can suppress artefacts not addressed by barcoding [29].
Figure 3.7: SmDeep: three HCT116 mutations identified through deep sequencing of a 5 gene panel across dilutions from 100% to $10^{-4}$%. Every base at each position was called to illustrate the frequency of background errors present. Top panel compares limit of detection with standard uncollapsed reads, middle panel represents single strand consensus sequences, and bottom panel highlights duplex consensus sequences.
Figure 3.8: LgMid: Five MOLM13 mutations identified through sequencing to moderate depth across dilutions from 100% to 0.04%. Every base at each position was called to determine frequency of known mutations and as well to illustrate the frequency of background errors. Each row corresponds to an error suppression method and each column represents a different mutation.
Figure 3.9: Summary trends of SmDeep mutations across different error suppression strategies. Comparison of detection limit using a) allele frequency and b) reads supporting each variant.
Figure 3.10: Summary trends of LgMid mutations across different error suppression strategies. Comparison of detection limit using a) allele frequency and b) reads supporting each variant.
Chapter 4

Discussion

4.1 Conclusions

Here, we proposed two strategies to improve molecular characterization and error suppression for detection of low-allele-frequency variants across two applications: 1) enhance molecular distinction using alignment information for deep sequencing (SmDeep), and 2) expand error suppression to singletons for moderate depth sequencing (LgMid).

With deep sequencing, we demonstrated the utility of alignment information (i.e. CIGAR strings) to increased identification of singletons by 33%. This improved quantification of molecular diversity for our small deeply sequenced panel and enabled additional error suppression in SSCS up to 70%. This has potential of improving distinction of rare fragments, such as cfDNA with tight size distributions where genome mapping and barcodes alone may be insufficient. Deep sequencing is required for consensus formation with traditional UMI methods as multiple reads are needed for error suppression. We established a novel approach for consensus construction enabling error correction within single reads. Utilizing the duplex methodology of aggregating complementary strands, we rescued singletons using their matching SSCS or singleton strand to remove technical artefacts. This expands the pool of reads for downstream consensus formation with up to 25% of singletons and improved efficiency rates for SSCS and DCS development. Interestingly, rescued singletons outperformed conventional SSCSs achieving similar error rates as DCS at $1 \times 10^{-4}$% error per target base. Upon further examination of error profiles, singleton rescue displayed similar substitution error features as DCS. Notably, corrected singletons successfully eliminated oxidative damage, which SSCS was incapable of despite having multiple read support. This is a phenomenon that can only be addressed through double-strand artefact suppression.

Through down-sampling of SmDeep, we mapped the trajectory of consensus formation with increasing depth of coverage. With rise in coverage, we observed a decrease in the number of reads required to construct each consensus. SSCS efficiency decreased upon saturating single-strand molecules at 8,000x coverage, while DCS saturation was not observed until 16,000x coverage. At the maximum efficiency rate, DCS was constructed from 3% of molecules in the library. This highlighted the inefficiencies with conventional
duplex barcoding as previously reported by other groups [71, 29]. With the incorporation of singleton rescue, we achieved DCS saturation between 4,000x and 8,000x coverage with maximum DCS efficiency doubling the conventional duplex method. Furthermore, inclusion of singleton rescue maximized theoretical DCS recovery regardless of sequencing depth. These observations can provide a reference to guide future sequencing experiments utilizing UMIs for optimal consensus establishment. Our method offers a lower limit of detection than conventional calling algorithms reaching 0.1% with deep sequencing. In addition, singleton rescue improved variant detection limit in DCS from 1% to 0.2% allele frequency with moderate depth sequencing. Together, our findings suggest UMI enabled error suppression can achieve lower limits of detection and has potential of improving sensitive detection of low-allele-frequency variants for clinical applications.

4.2 Potential impact and applications

Singleton rescue offers an error suppression strategy to address barriers for low-allele-frequency variant detection. This has potential applications for circulating tumour DNA detection in liquid biopsies. As sample input is often limited and signal is diluted by cfDNA from non-malignant cells, molecular recovery and suppression of background errors are critical for ctDNA analysis. This study highlights an improved strategy to depict molecular complexity and eliminate artefactual errors. The results emphasize the capability of mutation calling with DCS at a lower detection limit of 0.2%. This has implications for identifying ctDNA at frequencies below the current threshold set by background levels of technical artefacts. Furthermore, in silico assessment of molecular recovery across defined coverage intervals allowed us to quantify the frequency of unique molecules per target position. This distribution can be modeled to guide target panel design and depth of coverage for future liquid-biopsy sequencing experiments.

As much of the morbidity and mortality associated with cancer is related to late diagnosis of disease, earlier detection could improve timing and efficacy of interventions. ctDNA may be used to evaluate the onset and progression of disease. Although UMIs has been employed in ctDNA analysis for earlier diagnosis, identifying somatic alterations remains a major challenge without knowledge of matched tumour genotypes. As ctDNA is limited at early stages of disease, sensitivity may be improved with increased molecular recovery using complex identifiers reflective of individual sequence properties. In addition, singleton rescue may further increase sensitivity through error suppression of additional molecules. These refinements may improve screening of high risk individuals with family histories of carrying genetic mutations, paving the path for potential
diagnostics of disease in the general population.

While advancements are needed to enhance applications of liquid biopsies for screening, ctDNA has shown promise with detection of residual disease [26, 36]. Following surgery or treatments with curative intent, ctDNA analysis has demonstrated earlier identification of recurrence than current standard of care imaging [38]. With our sensitive method for single molecule error suppression, we hope to facilitate monitoring and enable earlier detection of disease. Currently, we have ongoing projects using UMI for detection of minimal residual disease in ctDNA of multiple myeloma patients, negating bone marrow biopsies [40]. Additionally, this technique has implications for monitoring relapse in patients post-surgery or transplantation. Particular impact is likely in diseases such as liver cancer where recurrence rates can be as high as 70% [80]. Overall, serial monitoring with ctDNA has relevance throughout the patient life cycle from screening to detection of relapse. Across many cancers, liquid biopsies can be exploited to improve clinical care and enable personalized treatment decisions.

4.3 Future directions

In this study, we observed a lack of molecular diversity within deeply profiled targeted panels. While we sequenced to 186,000x depth per target position in one experiment, we found that only 8,000x coverage was sufficient to generate SSCS from 27% of molecules in the library. We observed molecular saturation at 8,000x coverage for SSCS and 16,000x coverage for DCS. As higher depth of sequencing is required, DCS saturation is unlikely achievable with larger panels without significant increases in costs. Therefore, singleton rescue is needed to reduce the number of reads required to construct DCSs. In addition, singleton rescue maximized overall DCS recovery regardless of sequencing coverage across 9 intervals between 500x to 128,000x reads per target position. Together, our results support inclusion of singleton error suppression for all future UMI experiments as it is fundamental for optimizing DCS sequencing.

A major challenge remains with alteration of molecular barcodes caused by polymerase and sequencer errors. These technical artefacts modify UMIs causing mis-categorization of reads and inaccurate quantification of molecules. While deeper coverage profiling is often suggested as a solution to improve molecular recovery and sensitivity of detection, increased depth also results in a linear increase in UMI errors [81]. Conventional duplex strategies require deep sequencing for DCS recovery, whereas our improved strategy of singleton rescue reduces the necessity for high depth of coverage. Rather, our survey of molecular abundance across a broad range of coverage intervals urges against
over sequencing for economical reasons and avoidance of artificial UMIs generated with deep sequencing. Overall, this work will enable precise modelling of depth and expected molecular complexity to guide the design of future panels and targeted DNA sequencing studies.
Bibliography


