5-2016

Investigating Metastatic Lineage In Colorectal Cancer By Single Cell Dna Sequencing

Marco Leung

INVESTIGATING METASTATIC LINEAGE IN COLORECTAL CANCER BY SINGLE CELL DNA SEQUENCING

by

Marco Lokyin Leung, BS

APPROVED:

_____
Nicholas E. Navin, PhD
Supervisory Committee Chair

_____
Elsa R. Flores, PhD

_____
E. Scott Kopetz, MD, PhD

_____
Ralf Krahe, PhD

_____
Kenneth Y. Tsai, MD, PhD

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

INVESTIGATING METASTATIC LINEAGE IN COLORECTAL CANCER BY SINGLE CELL DNA SEQUENCING

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
And
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Marco Lokyin Leung, BS

Houston, Texas

May, 2016

To my parents, Timothy Leung and Minnie Yick

Acknowledgments

This dissertation would not be possible with the support of many people in and outside of the laboratory.

I am deeply indebt to my mentor, Dr. Nicholas Navin, for giving me the opportunity to work in his lab. He has provided me countless hours of mentoring, training and advice on my PhD projects and scientific career. Under his mentorship, I was able to expand my skillset and learn to critically analyze genomic data. He has supported my career aspiration since day one, and funded all my scientific meetings and bioinformatics trainings to achieve my goals. I am thankful to have a PhD advisor that provided me the resources, time and freedom to grow as a scientist.

I thank my current and past advisory committee members (Drs. Ken Chen, Elsa Flores, Andrew Gladden, Marilyn Li, Scott Kopetz, Ralf Krahe and Kenneth Tsai) for taking their time and offering invaluable advice for my projects. I would like to particularly thank Drs. Elsa Flores and Kenneth Tsai. I spent my last year of my undergraduate study in the Flores lab, and after graduation, worked a year as a research assistant in the Tsai lab. It was during my time in these two labs that led me to the decision to pursue my graduate studies. I am grateful for Drs. Flores and Tsai's time for their ongoing support in my advisory committee, and for writing me countless recommendation letters for my scholarships, fellowships and postdoctoral applications.

I am grateful for all the members in the Navin lab. The Navin lab has expanded over the past five years and has become a truly collaborative

environment, where each member provides his or her expertise to those in need. I enjoyed the healthy discussion and insightful critiques during lab meetings. I thank Dr. Yong (Tony) Wang for teaching me bioinformatics during my first two years of PhD and for providing invaluable scripts that made my analysis easier and faster. I thank Jill Waters for her experimental help during the early phase of my projects. I thank Charissa Kim for her astute advice on my projects and career prospects during coffee breaks, and Anna Casasent for her assistance in my clustering scripts in R. I thank Dr. Emi Sei and Pei Ching (Tessa) Tsai for always being there and lending their hands whenever I needed help the most. I thank Dr. Ruli Gao, Alexander Davis and Naveen Ramesh for teaching me the biostatistics and analytical tools for my projects. I also thank my summer students, Jerry Jiang, Jose Acevado Echevarria, and Jonathan Chen, for helping me with my experiments.

I feel fortunate to have been a part of the Genes and Development Graduate Program family. I thank Elisabeth Lindheim for her moral support during my studies and her help on my layman abstracts for many of my scholarship applications. I thank Dr. Deepavali Chakravarti for being the role model of a perfect graduate student and for her mentoring when I was an undergraduate student. I thank Dr. Blake Palculict for his advice and help for my ABMGG fellowship applications. I thank all the G&D students whom I have worked with during the annual retreats' entertainment hour, which was one of the highlights of my PhD studies.

I thank our neighboring labs, the Huff lab and the Lozano lab, particularly Dr. Amanda Wasylishen, for their experimental and reagent supports. I thank Erika Thompson, Dr. Louis Ramagli, Dr. Hongli Tang and Kanhav Khanna in the Sequencing and Microarray Laboratory for their speedy sequencing services that made this dissertation possible. I also thank Wendy Schober and Nalini Patel in the Flow Cytometry Laboratory for their flexibility to always let me squeeze in a flow-sorting appointment at the last minute. I am appreciative of the GSBS staffs, particularly Tracey Barnett and Lily D'Agnostino, for being a great support system for students, as well as providing unlimited Keurig coffee supply to keep me awake and working.

Finally, I would like to my parents, Timothy Leung and Minnie Yick, and my brother, Victor Leung, for giving me their loving supports over the years. They have always been encouraging and tolerable in my career choices and personal demands. Last but not least, I feel extremely fortunate to know Edward Gonzalez and Katherine Dextraze during my first year of graduate school. I thank Edward Gonzalez for giving me the unconditional support in the form of lunches from China Town, and Katherine Dextraze for being the best friend/landlord anyone can ask for.

INVESTIGATING METASTATIC LINEAGE IN COLORECTAL CANCER BY SINGLE CELL DNA SEQUENCING

Marco Lokyin Leung, B.S.

Advisory Professor: Nicholas E Navin, Ph.D.

Metastasis is the primary cause of human cancer deaths. Patients with metastatic colorectal cancer (mCRC) show only an 11% 5-year survival rate, compared to those without local or distant metastases (92% 5-year survival rate). Understanding the CRC tumor evolution may provide valuable insights on how to improve treatment in patients with mCRC. However, the genomic basis of metastasis has been difficult to study, in part due to the extensive intratumor heterogeneity at both the primary and metastatic tumor sites, and the low frequency of subclones with metastatic potential. Previous studies have applied conventional bulk next-generation sequencing (NGS) methods, which have limited ability to resolve intratumor heterogeneity.

To address this problem, we have developed a highly-multiplexed single cell DNA sequencing method that combines flow-sorting of single nuclei, multiple-displacement-amplification using Φ29 polymerase, low-input library preparation, library barcoding, targeted capture and NGS to generate high-coverage data from single cells. We validate this method by generating high coverage sequencing data from single human cells, with low allelic dropout and high detection efficiencies for single nucleotide variants.

Using this method, we sequenced 186 single cells from primary tumor and liver metastases from two mCRC patients to delineate the clonal architecture of the tumor and reconstruct their phylogenetic lineages. We also performed exome sequencing on the bulk tumor tissues. Our data identified a large number of nonsynonymous mutations that evolved in the root node during the earliest stage of primary tumor evolution and were maintained in all single cells during the clonal expansion of the tumor mass. We also identified a small number of mutations that were specific to the liver metastases, which are likely to play an important role in metastatic dissemination. Furthermore, we found three diploid cells with only *APC* mutations in CO5, which may represent the progenitor clones that gave rise to the primary and metastatic tumors. Using the single cell data, we construct phylogenetic trees, which revealed branched evolution in metastasis. Our data suggest that both mCRC patients are consistent with the late-dissemination model, in which the primary tumors evolved for a long period of time prior to the dissemination of clones to distant organ sites.

In summary, we have developed novel methods for single cell DNA sequencing, and applied these methods to gain unprecedented understanding of clonal evolution during metastasis in colorectal cancer.

Table of Contents

List of Figures

Chapter 1 - Introduction

Chapter 2 - Methods and Materials

Chapter 3 - NUC-Seq - Single Cell Whole Genome Sequencing

Chapter 4 - SNES: Single Nucleus Exome Sequencing

List of Tables

CHAPTER ONE – INTRODUCTION

Chapter 1 - Introduction

1.1 Colorectal Cancer

Colorectal cancer (CRC) is the third most common cancer among men and women in the United States with a lifetime risk of 1 in 20 (5%).[1] It is estimated that there are more than 130,000 new CRC cases for 2015.[1] The risk of CRC increases with age. Over 90% of new cases are diagnosed in people over 50 years old.[2] This is 15 times higher than people who are between 20 to 49 years old.[2] Men are more likely to develop CRC than women (57.2% vs. 42.5%), however it remains unclear why there is a disparity in CRC incidence between the two genders.[3] The pathogenesis of CRC is a long process, which usually takes more than a decade to develop.[4] CRC first develops as a polyp from the lining of the colon or rectum and progresses to adenocarcinoma. Per The American Cancer Society's recommendation, men and women should have colonoscopy as an early CRC screening beginning at the age of 50.[5] This practice has decreased the CRC incidence by 3.0% for men and 2.3% for women per year since the 1980s.[6] Colonoscopy was estimated to have prevented two-third of cancer deaths from the descending (left-side) colon.[7]

1.1.1 Chromosomal Instability and Microsatellite Instability

Using genome instability as markers, CRC can be categorized into chromosomal instability (CIN) and microsatellite instability (MSI).[8] CIN refers to the aberrant insertion and deletion of part or full chromosomes leading to aneuploidy; MSI refers to unstable DNA repeat length caused by a defect in the DNA repair pathway.[8] About 85% of CRC patients are CIN and 15% are MSI.[9] Although it

was thought that CIN and MSI are mutually exclusive, it was found that a portion of CSC cases present both.[10] MSI CRC tumors harbor higher number of mutations than CIN CRC.[11] This will be further discussed in a later section. (**1.4.1 Sequencing Studies and Colorectal Cancer**)

1.1.2 Staging of Colorectal Cancer

According to the American Cancer Society, CRC is categorized into different stages: stage 0 – also called *carcinoma in situ*, cancer has not grown beyond the inner mucosal colon or rectum layer; stage I – cancer has grown into the submucosa; stage II – cancer may have spread through the wall of the colon or rectum into nearby tissue or organs; stage III – cancer has spread to nearby lymph nodes; stage IV – cancer has spread to one or more distant organs.[12] Therapeutic treatments are usually determined by the locality and staging of the tumor.[13] For stage 0 and I CRC, surgery during colonoscopy is the standard of care to remove the section of colon or rectum that has the polyp.[13] For stage II CRC, surgical resection is still the standard treatment but chemotherapy may be necessary if there is a high risk of cancer coming back.[13] For stage III CRC, surgery with adjuvant chemotherapy is the standard of treatment.[13] For stage IV CRC, treatments are varied depending on the patients. If there are only few metastases, surgery will be performed to remove all the tumors. If there are too many metastases for surgery, chemotherapy will be given and then surgery may be tried.[13]

1.2 Models of Metastasis

While the overall 5-year survival rate for CRC has improved over the last 20 years, the survival rate for patients with stage IV CRC is only at 11-12%.[14] This grim outlook for CRC metastasis is consistent with other cancers, in which most metastases have poorer survival rates than primary cancers.[2] It was estimated that 90% of human cancer death is caused by metastases.[15] For CRC, the most common site of metastasis is the liver. Therefore, understanding the process of which CRC metastasize is critical to decreasing the mortality rate. Below, we describe the several general models of metastasis that have been proposed to describe how tumor cells disseminate from the primary tumor to distant organs.

1.2.1 Late Dissemination Model

The late dissemination model, also called the linear progression model or the classical model, is one of the earliest models to be proposed describing tumorigenesis. This model posits that cancer cells sequentially accumulate mutations and survive multiple rounds of clonal selections to form the primary tumor, and eventually metastasize after having acquired mutations that confer a metastatic phenotype.[16,17] The metastasis, in theory, is the most malignant cells of the primary tumors that can survive the circulation and seed at distant organ sites. This model implies that it takes years, sometimes decades, to form the primary malignant tumor, whereas metastases only need months to form.[16] In other words, the tumor has evolved for a long time before cells diverge and form metastasis in a short period of time; this leads to a similar mutational profile shared between the primary and metastasis. (Figure 1)

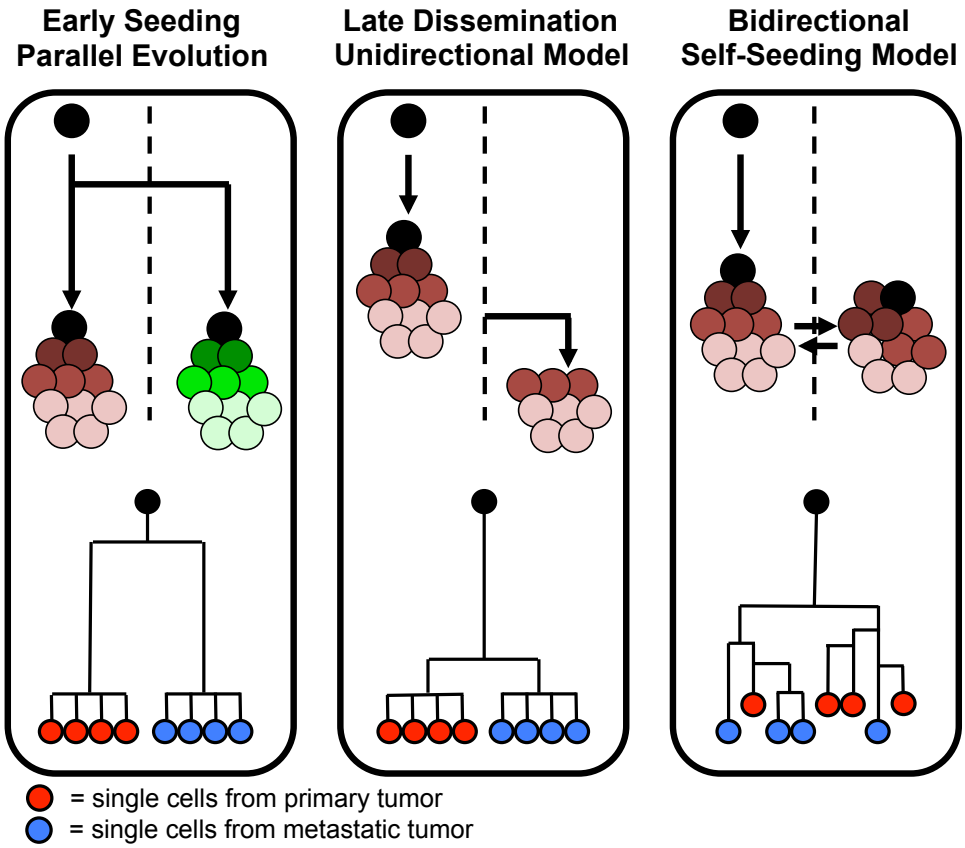

Figure 1 - Model of Metastasis

The early seeding model posits that cells disseminate from the primary tumor during early tumor development, leading to parallel evolution of two distinct genomic profiles. The late dissemination model posits that cells leave the primary tumor during late stage of tumor development, thus retaining the majority of genomic signature. The self-seeding model posits that tumor cells travel bi-directionally between the primary and metastatic sites, leading to the intermixing lineage of tumor cells.

1.2.2 Early Dissemination Model

The early dissemination model, also called the parallel progression model, is also another model proposed to describe the metastatic process. This model suggest that cancer cells disseminate from the primary tumor very early on during tumorigenesis, and acquire metastatic somatic mutations in parallel, leading to independent lineages with different mutational profiles. (Figure 1) It was suggested that, although the tumor cells disseminate early on during tumorigenesis, the metastasis might not arise at a similar time as the primary tumor. This may be due to cancer dormancy, in which micrometastases form and enter into a quiescent state.[18,19] The metastases then reenter into a proliferative state triggered by different factors.[19]

The early dissemination model may also explain a clinical cancer syndrome, called cancer of unknown primary (CUP).[20] CUP may be due to the early dissemination of tumor cells, which acquire somatic mutations at the remote site and become malignant before the primary tumor. The late dissemination model, in contrast, cannot explain this phenomenon.

1.2.3 Self-Seeding Model

The self-seeding model was proposed recently and it posited that circulating tumor cells (CTCs) travel from metastasis back to the site of origin.[21] This model suggests that metastasis is a bidirectional, rather than unidirectional, process.[21] (Figure 1) It was shown in mice, that IL-6 and IL-8 act as CTCs attractants to recruit myeloid cells into the stroma.[21] In other words, these CTCs can promote tumor growth in primary and metastatic sites by altering the microenvironments.[22]

Therefore, it is reasonable to warrant further investigation and potentially target these CTCs to stop the self-seeding process. However, this model has only been shown in mouse models and it has been difficult to examine the self-seeding model in human with current tools and technologies.

1.3 Intratumor Heterogeneity

Intratumor heterogeneity refers to the morphological and molecular differences within individual tumors and has been reported in most solid cancer types.[23-26] These differences may include histology, genotype, gene expression and proliferative potential.[27] For example, using fluorescence *in situ* hybridization (FISH), the copy numbers of specific chromosome loci can be quantified across many cells, determining the clonality within a tumor.[28] Similarly, using Giemsa staining on cells in which chromosomes are in metaphase, larger scale of chromosome aberrations can be observed.[28] Although intratumor heterogeneity has been observed and studied as early as 1800s, the origin of intratumor heterogeneity during tumor evolution remains debatable and several competing hypotheses have been proposed.[27,29] One of the earliest models to explain intratumor heterogeneity is the clonal evolution model, originally proposed by Peter Nowell, who was the first to propose an evolutionary model for tumor growth.[30] (Figure 2) Monoclonal evolution model suggests that intratumor heterogeneity occurs early in tumor progression, leading to one dominant population of tumor cells that expands to form the mass of the tumor, whereas polyclonal evolution model proposes that there are multiple clones co-existing, that have the proliferative potential and expand and form the tumor mass.[28]

Another model is the cancer stem cell (CSC) hypothesis. This model posits that a minor population of tumor initiating cells has self-renewing properties allowing them to proliferate indefinitely, and gives rise to the majority populations of tumor cells that have limited replication potential.[28,31] (Figure 2) CSCs can be identified with different cell surface markers. However, it is difficult to study the genomes and transcriptomes of CSCs due to the rare frequencies of CSCs within tumors, which are often well below one percent. The self-seeding model is another hypothesis that can also be used to explain intratumor heterogeneity. (Figure 2) When tumor cells from metastasis leave and re-seed at the primary site, the metastatic clone would aggregate on the outer region of primary tumor.

Intratumor heterogeneity has complicated the diagnosis and treatment of cancer patients. If a tumor has multiple clones in the tumor mass, it is necessary to sample from multiple spatial regions to detect mutations in clinical assays. Furthermore, if tumor contain subclones with different sensitivity for a specific therapeutic drug, administering treatments will have to be adjusted depending on the composition of less-sensitive subclones. Therefore, it is critical to detect intratumor heterogeneity with high sensitivity in the clinics. There are different methods and technologies that can be used to measure intratumor heterogeneity, including cell surface markers, immunohistochemistry, fluorescent *in-situ* hybridization (FISH) and genotyping of specific mutations. Next-generation

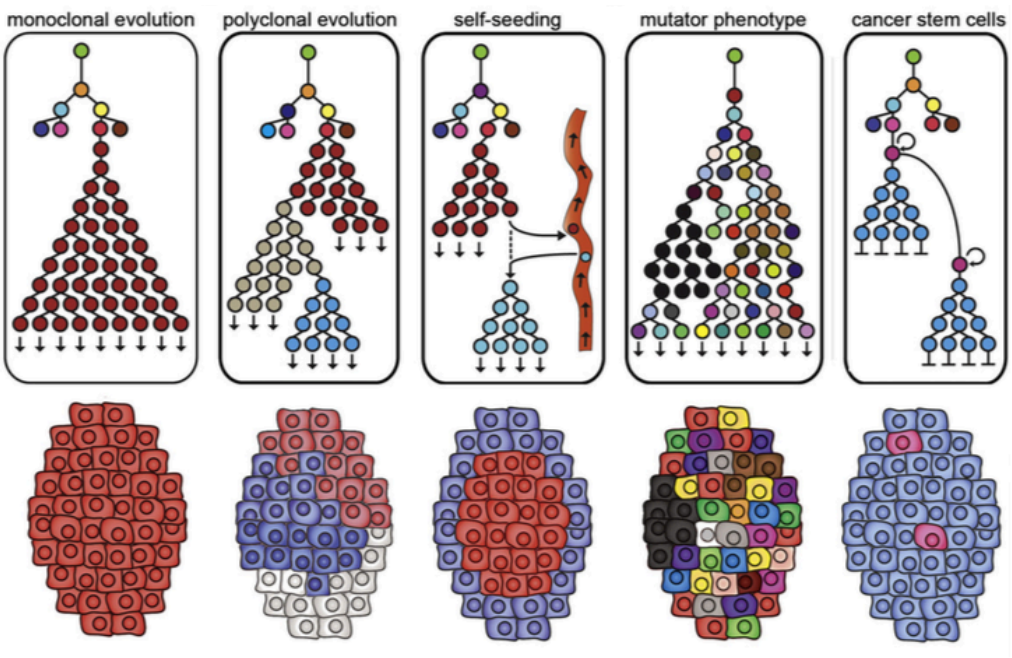

Figure 2 - Tumor Progression Model

This figure illustrates the forms of heterogeneity caused by the differences in how tumors progress. The monoclonal evolution model suggests that one clone expands and forms the tumor mass, whereas the polyclonal evolution model suggests that there are multiple clones that have the capability to contribute to the tumor mass. The self-seeding model posits that tumor cells can travel bi-directionally between the primary and metastatic sites. The mutator phenotype can generate a tumor with many diverse clones. The self-seeding model posits that a minor population of cells has the self-renewing property, which gives rise to the majority of cells.

(Modified and reproduced from Navin, N. E. & Hicks, J. Tracing the tumor lineage. *Molecular oncology* **4**, 267-283, doi:10.1016/j.molonc.2010.04.010 (2010) with permission from Elsevier Limited.)

sequencing (NGS) has been used to measure genomic intratumor heterogeneity. Below, we will discuss NGS and its roles in resolving intratumor heterogeneity in more detail.

1.4 Next Generation Sequencing

Over the past 10 years, the advent of NGS has revolutionized the fields of biology by exponentially increasing the data output of genome sequencing and vastly decreasing the cost per base.[32] NGS uses massively parallel sequencing to analyze millions of DNA fragments are sequenced at the same time. This is a major improvement over Sanger Sequencing, which uses the chain-termination method on only one single-strand DNA template at a time. Human Genome Project was able to complete a draft of the human genome in 2003, partly due to the high throughput of NGS.

Among the several sequencing platforms, Illumina-sequencing is the most commonly used platform for NGS. Illumina-sequencing includes two major steps, library construction and sequencing. In short, genomic DNA is fragmented into ~250-500 base pairs (bp) and the ends of fragments are repaired from sticky ends to blunt ends and the 5' end is then adenylated. Adaptors are then ligated onto the both ends of DNA libraries and they are amplified during PCR. Libraries are hybridized onto the probes of the sequencing flow cells and generated into clusters of libraries. Using the sequencing-by-synthesis approach, when each specific base is added onto the libraries, a corresponding fluorescent label is excited and emits a specific wavelength. The Illumina machine captures the

image of the whole flowcell with all the labeled clustered and converts signals into text data.

1.4.1 Sequencing Studies and Colorectal Cancer

Before the advent of sequencing technologies, most genes mutated in CRC were discovered using linkage studies and cytogenetic studies, and this led to the identification of the frequently mutated genes, such as *APC*, *KRAS*, and *TP53*.[8,33-35] Later, Sanger sequencing discovered more genes that are associated with CRC.[36-39] However, exome- and genome-wide studies using Sanger sequencing is costly and time-consuming, making it unfeasible to use this approach for routine research studies.

Since 2011, NGS has been used to investigate CRC and the number of mutations detected in these patients has increased drastically. Early NGS studies of CRC used whole-genome sequencing to find that there are estimated 75 somatic rearrangement in each of 9 sequenced non-MSI tumors.[40] On average, they found there are approximately 5.9 mutations per Mb, with 79 non-synonymous mutations per sample, as well as recurrent *VTI1A*-*TCF7L2* fusion.[40] Another comprehensive NGS study was conducted by The Cancer Genome Atlas, in which they sequenced more than 200 CRC tumors and analyzed the DNA exome, copy number, epigenetics and transcriptomic aberrations.[11] From these data, they categorized CRC tumors with more than 12 mutations per Mb as hypermutated, and those with less than 8 as non-hypermutated, and they found genes that were frequently mutated in each category.[2] Another study was published using both exome and RNA sequencing on 70 CRC samples,[41] in

which they found recurrent fusion events, such as *RSPO2-EIF3E* and *RSPO3-PTPRK* in 10% of CRC.

These large-scale NGS studies provided detailed analyses of intertumor heterogeneity – the genomic differences between patients. However, these studies focused mainly on primary tumors and did not investigate mutations in matched metastatic tumor samples.

1.4.2 NGS and Intratumor Heterogeneity

NGS studies, such as TCGA, sequenced tumors to discover mutations that are frequent in hundreds of patients in many different cancers types. [42-58] (Table 1) These efforts led to better understanding in intertumor heterogeneity. However, due to the inadequate coverage depth and the limitation of conventional NGS technologies, intratumor heterogeneity is often difficult to measure from these dataset. To improve measurements of intratumor heterogeneity, different NGS approaches have been developed.

Previous studies have used deep sequencing methods to understand intratumor heterogeneity.[59-61] By sequencing tumors at very high coverage depths, subclonal mutations can be discovered. Early NGS study used Pyrosequencing on B-cell chronic lymphocytic leukemia (CLL), in which the authors detected rare *IGH* locus with frequencies as low as 0.02%.[23] However, conventional deep sequencing methods have inherent error rates (0.1-1%) that may lead technical artifacts. A recent deep sequencing technology, called duplex sequencing, was developed to address this problem.[62] By using a random tag in the PCR primers, only variants that are presented multiple duplicates of the same

Year	Tumor Type	Tumors Studied	Journal	Reference
2008	Glioblastoma	206	Nature	42
2011	Serous Ovarian Carcinoma	489	Nature	43
2012	Colorectal carcinoma	276	Nature	44
2012	Squamous Cell Lung Carcinoma	178	Nature	45
2012	Breast Cancer	510	Nature	46
2013	Acute Myeloid Leukemia	200	NJEM	47
2013	Endometrial Carcinoma	373	Nature	48
2013	Clear Cell Renal Cell Carcinoma	417	Nature	49
2014	Urothelial Bladder Carcinoma	131	Nature	50
2014	Lung Adenocarcinoma	230	Nature	51
2014	Gastric Adenocarcinoma	295	Nature	52
2014	Papillary Thyroid Carcinoma	496	Cell	53
2014	Head and Neck Squamous Cell Car.	279	Nature	54
2015	Diffuse Lower-Grade Gliomas	293	NEJM	55
2015	Cutaneous Melanoma	331	Cell	56
2015	Papillary Renal-Cell Carcinoma	161	NEJM	57
2015	Prostate Cancer	333	Cell	58

Table 1 - List of Studies Published by The Cancer Genome Atlas

The table provides a partial list of cancer types sequenced by TCGA. The number of tumors sequenced in each study ranges from 131 to 510.

molecular tags are scored.[62] This authors showed that this approach led to a false-positive rate of less than 1 in $1x10^9$ nucleotides.[62]

Despite the improvement of accuracy of deep sequencing, this method is unable to resolve spatial heterogeneity. An alternative approach for resolving spatial heterogeneity involves sampling and sequencing multiple macroscopic regions of a tumor to infer intratumor heterogeneity. For example, primary renal carcinomas and matched metastatic sites were macro-dissected and exome-sequenced.[63] The authors found that over 60% of somatic mutation are not present across all spatial regions of the tumor mass. The same approach was also demonstrated in lung adenocarcinoma and in non-small cell lung cancer by two separate groups.[64,65] Another study used array comparative genome hybridization (aCGH) to study copy number profiling in multiple section in breast cancers and found that some breast tumors contain multiple tumor subpopulations.[66] These studies underscore the importance and significance of understanding intratumor heterogeneity using genomic technologies. However, depending on the size of the tumor dissection, it is still difficult to detect mutations that occur at low frequencies using spatial sequencing (<1%).

1.5 Single Cell Sequencing

Single cell sequencing (SCS) offers an alternative approach to study intratumor heterogeneity.[67] By sequencing many cells across multiple sections of a tumor, spatial heterogeneity can be studied and the mutations present in rare subclones can be detected. SCS can achieve the goals of both spatial sequencing and deep sequencing. The first single cell RNA sequencing study was published in

2009.[68] After performing reverse-transcription of mRNA, cDNA is amplified by PCR and constructed for sequencing libraries. The authors showed that single cell RNA sequencing was more sensitive compared to other platforms, such as Affymetrix microarrays. Since then, further development of single RNA sequencing has increased the throughput and decreased the cost, enabling thousands of single cells to be sequenced at one time (for example, Drop-Seq).[69-71] On the other hand, the development of single cell DNA sequencing methods has been more challenging. This is partly because there are thousands of copies of each RNA molecule, whereas there are only two copies of each DNA molecule in each single cell.[67] This limited amount of input material gives rise to the technical errors, such as false positive and allelic dropout.[67]

1.5.1 Previous Published Single Cell DNA Sequencing Methods

The first single cell DNA sequencing method, called Single Nucleus Sequencing (SNS), was published in 2011 to study the intratumor heterogeneity in breast cancer.[72] To amplify the single cell genome, the authors used degenerate oligonucleotide primed PCR (DOP-PCR), which primers with semi-random nucleotides are used to generate genome-wide DNA with minimal bias.[73,74] Single cell libraries are then sequenced at sparse coverage depth. The genome is divided into small bins (50,000 bp) and reads are counted within each bin to infer copy number status. This approach was used to sequence two high-grade, triple-negative (ER-, PR- and HER2-) ductal carcinomas and a paired liver metastasis.[72] By sequencing 100 nuclei from each patient, they found that there were 3 distinct clonal subpopulations in a polygenomic tumor, whereas there was

only one population in the monogenomic tumor and its liver metastasis.[72] This study was the first to demonstrate using single cell DNA sequencing to infer tumor progression. However, this method is not sufficient to score single nucleotide variants due to the low coverage depth (0.5x).

Two single cell DNA sequencing studies were later published simultaneously.[75,76] These two studies performed exome sequenced of single cells of clear cell renal cell carcinoma and myeloproliferative neoplasm.[75,76] Using the multiple-displacement-amplification (MDA) approach, these studies used phi29 to amplify genomes and capture the exonic regions for sequencing. It was shown that phi29 polymerase has a superior performance with low false positive error rate of $1x10^{-7}$ compared to *bst* polymerase ($1x10^{-5}$), which does not have proofreading capability.[67,77,78]

Another single cell DNA method is called MALBAC – multiple annealing- and looping-based amplification cycles.[79] This method uses specific primers to loop both end of amplicons, thus preventing copied DNA to amplify exponentially. In theory, WGA reaction only amplifies the original DNA template, thus decreasing technical errors from amplifying. [79] This method can be used for both copy number and single nucleotide analysis, however it produces high false positive error rates,[67] possibly due to the use of *bst* polymerase instead of phi29 polymerase.

1.5.2 Challenges in Single Cell DNA Sequencing

Despite the recent advances in single cell sequencing in the last five years, there are still major technical challenges.[67] Because of the low amount of DNA input in

single cells (6 picograms), amplification using a DNA polymerase is needed to generate sufficient amount of input material for library construction. However, DNA polymerases are prone to introducing artificial errors during the amplification step, thus generating high false positive error rates. False positive errors are problematic for the analysis of mutations in single cells, as they are difficult to distinguish from real biological events.

Another major SCS technical challenge is high allelic dropout rate. Allelic dropout refers to the detection of only one allele of two heterozygous alleles in single cells. Because WGA polymerases have allelic preference during the genome amplification step, it is common to have uneven amplification of both alleles. If the variant alleles are completely dropped out, it leads to false negative during analysis and mutations are missed.

Low throughput is another problem for single cell DNA sequencing. For single cell RNA sequencing, various methods have been developed to multiplexed thousands of single cells into one sequencing lane. For example, Drop-Seq uses a microdroplet system to simultaneously perform reverse transcription of mRNA and amplification and cDNA of up to thousands of single cells.[69] This is possible because of the polyA tail in which an adaptor with a cell ID barcodes can be easily ligated using synthetic beads. Furthermore, to evaluate gene expression of single cells, sequencing reads are counted for each gene, and high coverage depth is not required.

Unlike RNA sequencing, multiplexing high number of single cell for DNA sequencing remains difficult and technically challenging. Because of the

polymerase used during the WGA step, long molecules of DNA are generated (>10 kilobases) and a sonication step is needed afterwards to construct libraries. Therefore, barcoding during WGA is not feasible. Furthermore, to generate enough sequencing reads to analyze variants at single nucleotide level, at least 30x of coverage depth is needed, thus multiplexing many cells is not feasible for genome or exome sequencing.

1.6 Dissertation Summary

This dissertation focuses on developing novel single cell DNA sequencing methods to study intratumor heterogeneity and metastatic evolution in colorectal cancer. This chapter has so far described the background and current challenges in single cell DNA sequencing, and knowledge gaps in CRC and metastasis. This dissertation aims to resolve the following problems.

1. Intratumor heterogeneity is difficult to study using standard NGS technologies since they are limited to reporting a bulk admixture of genomes in tissue samples.
2. Genome evolution during CRC has been difficult to study due to intratumor heterogeneity at the primary and metastatic tumor sites.
3. Previous SCS methods have limited coverage breadth, and are unable to detect mutation at base pair resolution.
4. Current SCS methods are challenged by high false positive rates and high allelic dropout rates.
5. Current SCS DNA methods have low throughput and high costs associated with performing the experiments.

These chapters in this dissertation are written in the chronological order of how we developed different SCS DNA methods over the past five years. This dissertation aligns with the continuing efforts of our laboratory to develop and improve single cell DNA sequencing methods for cancer research. In Chapter 3, we present the development of the first single cell whole-genome sequencing method, NUC-Seq, to understand the intratumor heterogeneity in ER^+ and triple negative breast tumors. In Chapter 4, we further developed our method for exome sequencing of single cell and discuss the associated methodologies and error rates. In Chapter 5, we increase the throughput of single cell sequencing by multiplexing 96 single cells into one sequencing lane. In Chapter 6, we used SCS to study models of metastasis and intratumor heterogeneity in two CRC patients.

CHAPTER TWO – METHODS AND MATERIALS

Chapter 2 – Methods and Materials

Content of this chapter is based on the following publications:

Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. Clonal Evolution in Breast Cancer Revealed by Single Cell Genome Sequencing. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600. PMID:25079324.

Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

2.1 Cell Lines and Human Tumor Samples

For the NUC-Seq study in Chapter 3, SK-BR-3 is a Her2 positive ($ER^-/PR^-/HER2^+$) breast cancer cell line. The estrogen receptor positive breast cancer (ERBC) and triple-negative breast tumor (TNBC) samples used in this study were obtained from the MD Anderson Cancer Center Breast Tissue Bank as frozen tumor specimens. Histopathology classified both breast tumors as invasive ductal carcinomas. The ERBC was reported to have mixed invasive lobular carcinoma. Both tumors were excised by lumpectomy before any chemotherapy or radiation therapy. The ERBC tumor grade was scored as Nottingham histological grade 2, whereas the TNBC tumor was scored as grade 3. Receptor staining showed that the ER tumor was positive for estrogen receptor (80%), positive for progesterone receptor (90%) and negative for the HER2 receptor (FISH HER2/CEP17, ratio 1.1). The TNBC was negative for estrogen receptor (2%), negative for progesterone receptor (3%) and negative for the Her2 receptor (FISH HER2/CEP17, ratio 1.3). This study was approved by the Internal Review Board (IRB) at MD Anderson Cancer Center.

For the SNES study in Chapter 4, SKN2 is a human fibroblast cell line that was obtained from the Cold Spring Harbor Laboratory (Dr. Michael Wigler). SKN2 was cultured using Dulbecco's Modified Eagle Medium with 10% fetal bovine serum, penicillin/streptomycin and L-glutamine.

For highly-multiplexed single cell DNA sequencing method in Chapter 5, we use MDA-MB-231, which is a triple-negative breast cancer cell line. This cell line uses the same media condition as SKN2.

For the CRC study in Chapter 7, tumor samples from two CRC patients (CO5 and CO7) were collected from MD Anderson Tumor bank through collaboration with Drs. Scott Kopetz (Department of Gastrointestinal Medical Oncology) and Dipen Maru (Department of Pathology). CO5 is a 77-year-old CRC patient with invasive moderately to poorly differentiated adenocarcinoma with liver metastasis. At the time of surgery, the colon primary tumor size was 4.0 x 3.0 x 1.0 cm and the liver metastasis size was 4.1 x 2.3 x 2.0 cm. CO8 is a 64-year-old CRC patient with invasive moderately differentiated adenocarcinoma with liver and lung metastasis. The colon primary tumor size was 4.0 x 5.0 x 1.5cm and the liver metastasis size was 0.4 x 0.2 x 0.2 cm. However, we do not have the lung metastasis.

2.2 Single Cell Isolation

Nuclei of cell lines and frozen tumors were isolated using NST/DAPI buffer (800mL of NST (16mM NaCl, 10mM Tris base at pH 7.8, 1mM $CaCl_2$, 0.05% BSA, 0.2% Nonidet P-40)), 200mL of 106 mM $MgCl_2$, 10mg DAPI and 0.1% DNase-free RNase A. Cultured cells were trypsized and lysed directly in NST/DAPI buffer. Sectioned tumors were cut and minced using surgical blades in a Petri dish in NST/DAPI buffer in the dark. Samples were filtered through a 37-µm plastic mesh to a 5-mL polystyerene tube. Nuclei were then sorted using FACS Aria II (BD Biosciences) and single nuclei were deposited into individual wells on a 96-well plate. Single nuclei were gated from the G2/M distribution of cells.

2.3 Single Cell Genome Amplification

For the NUC-Seq project in Chapter 3, sorted single cells were amplified using REPLI-G UltraFast Mini Kit (Qiagen, #150035) per manufacturer's instructions with minor modifications. In short, we put lysis buffer in each well of the 96-well plate prior to flow-sorting. After flow-sorting, we centrifuge the plate and incubate at 65°C for 10 minutes. We incubate the single cell DNA with Φ29 polymerase at 30°C for 80 minutes. DNA was purified using QIAamp DNA Blood Mini kit (Qiagen, #51104) and quantified using the Qubit 2.0 Fluorometer (Invitrogen, Q32866).

For Chapter 4-7, the single cell amplification step uses in-house lysis buffer and amplification buffers. Make a 2:3 ratio of lysis buffer(200mM KOH, 50mM DTT):1xPBS solution. Load 3.5μL of solution into each well of a 96-well plate. After flow sorting, plate is centrifuged at 130g for 1 minute at room temperature. 1.5μL of neutralization buffer (900mM Tris-HCl, 300mM KCl, 200mM HCl) is added into each well and centrifuge. Amplification is performed using Φ29 polymerase (NEB, M0269L) with 500μM hexamers (with phosphorothioate modification at the last 2 bases) and 1mM dNTP (GE Healthcare, 28-4065-52). Final reaction volume is 50μL per well. Incubate at 30°C for 3 hours and 65°C for 3 minutes. Refer to Leung *et al* in *Nature Protocol* for detailed composition of each buffer.

2.4 Quality Control for Amplified Single Cell Genome

For WGA reaction, there are 22 individual qPCR reactions flanking a 200bp region of each autosome. (Table 2)

Name	Chromosome	Orientation	Primer Sequences	Product Size (bp)
chr1f	1	forward	TCCAAGCTCCAGTTCCAGAT	196
chr1R	1	reverse	TGCACTGAGACCTTCACAGG	
chr2f	2	forward	AGCGGGAGGGACTATTTCAC	200
chr2r	2	reverse	GGATCGTTCAAAGGGAACTG	
chr3f	3	forward	CCCTTGTACTGGCTCGTGTT	204
chr3r	3	reverse	CTTGCACATGAAGGTCTGGA	
chr4f	4	forward	GAGCATCTCTTGGCTCTGCT	210
chr4r	4	reverse	TTGGGAAAGCACAGATCCTT	
chr5f	5	forward	TTGCAGCTTTCCATTACGTG	208
chr5r	5	reverse	CCTTTTATGCCTCCAGCATC	
chr6f	6	forward	GAGGAGGGCAAGGAGAGAGT	202
chr6r	6	reverse	ACCCTCCAGTGTGCAAAAAC	
chr7f	7	forward	CTTCCTGCCATTCCACAAGT	210
chr7r	7	reverse	CCCACTTTCATGCCTCTGAT	
chr8f	8	forward	CTTCCCTGCCTTGCTCTCTA	207
chr8r	8	reverse	CGGGACATTTCAGCAATCTT	
chr9f	9	forward	CTGTGGAGCAGCTGTTTCTG	204
chr9r	9	reverse	GAATTCACAAAGCCCCAAGA	
chr10f	10	forward	CCCCTCATTCAAATCAGCAT	205
chr10r	10	reverse	CAGGCAAAAGCTGGAGTTTC	
chr11f	11	forward	AGCATCATCCAGCCCATTAC	210
chr11r	11	reverse	AAATCCCTGCAGAGCAGTGT	
chr12f	12	forward	ATCATGGAAATGCAGCCTCT	192
chr12r	12	reverse	AGAACCCAGCTGGAATGATG	
chr13f	13	forward	TGTTTCATGGAGTCCTGCTG	202
chr13r	13	reverse	GGAGGCAAGAACCAAACAAA	
chr14f	14	forward	AGCCAAGACGTACCCTCTCA	208
chr14r	14	reverse	TGCTTTACACCAATCCCACA	
chr15f	15	forward	TCAGCATGGGTTATGGGTTT	196
chr15r	15	reverse	CCCAGATGATGGAGAGGAAA	
chr16f	16	forward	GCCTGTGTTTGCTGATGAAA	193
chr16r	16	reverse	GGGCAACGACCGTACTTAAA	
chr17f	17	forward	TCCTGGGCTAGCCTTTTACA	199
chr17r	17	reverse	ATCGCTTGAGCACTGAAGGT	
chr18f	18	forward	AGACGAGCCTTTCTCTGTCG	203
chr18r	18	reverse	TCGAGACCATCCCCACTAAC	
chr19f	19	forward	TACTCAAAGCTGGCAGCAGA	194
chr19r	19	reverse	GAGCATGCCCAGGATACCTA	
chr20f	20	forward	CACCAGGGTCTTGATGGAGT	202
chr20r	20	reverse	AGCTCTGGGATCTGTGATGG	
chr21f	21	forward	TGGACAAATAAAGGCAATGG	190
chr21r	21	reverse	TCAGGCAACTTCTGGATGAA	
chr22f	22	forward	CTAGGATCCCGTGAAGGTCA	202
chr22r	22	reverse	AGGTAAGGGGACTCCTTGGT	

Table 2 - List of qPCR Primer Sequencing for Quality Control

For Chapter 3, each PCR reaction is set up using primers according to the table below. Perform PCR using the following conditions.

Temperature	Duration	
95°C	30s	
95°C	30s	30 cycles
60°C	60s	
68°C	60s	
68°C	5m	
4°C	hold	

For Chapter 4-6, each qPCR reaction is set up using primers according to the table below. Perform qPCR using the following conditions.

Cycle Numbers	**Denature**	**Anneal/Extension**
1	95°C, 3 minutes	
2-46	95°C, 20 seconds	60°C, 30 seconds

2.5 Library Preparation

For NUC-Seq study in Chapter 3, the WGA DNA was incubated with the Nextera transposome (Epicentre, Inc) to perform a tagmentation reaction in HMW buffer according to manufacturer's instruction. The libraries were purified using MinElute PCR purification kits (Qiagen, #28106), followed by 4 cycles of PCR. After PCR, the libraries were run on 2% agarose gels and size-selected in the 200-300 bp range (SK-BR-3) or 400-500 bp range (human tumors). The excised gel blocks were purified using MinElute purification columns. The size distribution

and concentration of the libraries were determined using the Bioanalyzer 2100 system (Agilent) using high sensitivity DNA microcapillary chips. The final concentration of the library was determined using qPCR with the KAPA library Quantification Kit (KAPA Biosystems, KK4835) and fluorescence was measured using the Qubit 2.0 system (Invitrogen, Q32866). For libraries that are prepared by ligation cloning, 100ng to 1000ng of DNA was acoustically sonicated to 300bp or 500bp using the Covaris Sonicator S220. Libraries were constructed using NEBNext DNA library Prep Master Mix Set for Illumina (New England Library, #F6040L) for end repair, 3' adenylation and ligation according to the manufacturer's instructions. MinElute PCR Purification Kit (Qiagen, #28006) is used for the purification step during library prep. Agarose electrophoresis is run for excision at 300bp to 400bp for size selection. We then perform 8 cycles of PCR following the manufacturer's instructions, using PE5/7 primers (Illumina Inc). Agencourt AMPure XP (Beckman Coulter, #A63881) was used for final purification. Final concentration was measured by qPCR using KAPA Library Quantification Kit (KAPA Biosystems, KK4835) and ABI PRISM real-time machine (Applied Biosystems 7900HT), as well as 210 Bioanalyzer (Agilent).

For Chapters 4 – 6, the WGA DNA is fragmented using Covaris Sonicator to size of 250 bp (peak incident power:157, duty factor:10%, Cycles per burst: 200, Treatment time: 130) and purified by Zymo DNA Clean & Concentrator Column Kit (Zymo, D4004) according to manufacturers instructions. Libraries are constructed using NEBNext end repair model (NEB, E6050L), dA-tailing module (NEB, E6053L) and quick ligation module (NEB, E6056L). The sequences for P5

adaptor and barcoded P7 adaptors are shown in Table 3. Adaptor-ligated libraries are amplified using NEBNext high-fidelity 2x PCR master mix (NEB, M0541L) with primers specific to adaptors (F: 5-AATGATACGGCGACCACCGAGATCTACAC-3 and R: 5-CAAGCAGAAGACGGCATACGAGAT-3)

For Chapter 6, single cell copy number profiling is performed for both colorectal patients. Cells are amplified using DOP-PCR according to the *Nature Protocol* paper by Baslan *et al*.[74] WGA DNA is run on agarose gel for quality control. Libraries are constructed for cells that passed quality control, following the *Nature Protocol* paper by Leung and Wang *et al.*

2.6 Exome/Targeted Capture

For Chapter 3 and 4, exome capture was performed on single cell sequencing libraries using the TruSeq Exome Enrichment Kit (Illumina, 15013230) following manufacturer's instructions with one modification: Nextera PCR primers (Epicentre) are used in place of the TruSeq PCR primers for library amplification. The capture platform targeted a 64Mb region including exons, promoters and UTRs. Final samples are purified using the AMPure XP beads (Beckman Coulter, A63881).

For Chapter 5, targeted capture was performed using Nimblegen SeqCap EZ Choice Library. The capture region covers 201 cancer-related genes defined in Chen *et al*.[80] For Chapter 6, targeted capture was designed to cover 1,000 genes. Targeted capture was done with 68-72-hour incubation. The procedure is performed according to manufacturer's instructions.

Adaptor with barcode	Sequence
1	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAAGACATCTCGTATGCCGTCTTCTGCTTG
2	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATGAATCATCTCGTATGCCGTCTTCTGCTTG
3	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTTAGCATCTCGTATGCCGTCTTCTGCTTG
4	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGACAGTGCATCTCGTATGCCGTCTTCTGCTTG
5	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAACAGGCATCTCGTATGCCGTCTTCTGCTTG
6	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTAAGGTCATCTCGTATGCCGTCTTCTGCTTG
7	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGGATTGCATCTCGTATGCCGTCTTCTGCTTG
8	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGACACACATCTCGTATGCCGTCTTCTGCTTG
9	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCTCTATCATCTCGTATGCCGTCTTCTGCTTG
10	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCTAATCCATCTCGTATGCCGTCTTCTGCTTG
11	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGACAACATCTCGTATGCCGTCTTCTGCTTG
12	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCATCCTCATCTCGTATGCCGTCTTCTGCTTG
13	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACCTTACATCTCGTATGCCGTCTTCTGCTTG
14	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAAGGAGCATCTCGTATGCCGTCTTCTGCTTG
15	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTGGCTCATCTCGTATGCCGTCTTCTGCTTG
16	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCATTCCATCTCGTATGCCGTCTTCTGCTTG
17	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAGCGACATCTCGTATGCCGTCTTCTGCTTG
18	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGGCTAACATCTCGTATGCCGTCTTCTGCTTG
19	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCCATGCATCTCGTATGCCGTCTTCTGCTTG
20	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCACCTCATCTCGTATGCCGTCTTCTGCTTG
21	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGATGTACATCTCGTATGCCGTCTTCTGCTTG
22	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGATTCATCTCGTATGCCGTCTTCTGCTTG
23	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACACGACCATCTCGTATGCCGTCTTCTGCTTG
24	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAATGTTGCATCTCGTATGCCGTCTTCTGCTTG
25	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAATCCGTCATCTCGTATGCCGTCTTCTGCTTG
26	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAGGACACATCTCGTATGCCGTCTTCTGCTTG
27	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAGAGATCATCTCGTATGCCGTCTTCTGCTTG
28	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACTCACCATCTCGTATGCCGTCTTCTGCTTG
29	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTCACGCAATCTCGTATGCCGTCTTCTGCTTG
30	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGGTGGTAATCTCGTATGCCGTCTTCTGCTTG
31	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGGCTTCAATCTCGTATGCCGTCTTCTGCTTG
32	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGGAACAAATCTCGTATGCCGTCTTCTGCTTG
33	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGAAGAGAATCTCGTATGCCGTCTTCTGCTTG
34	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCTTCACAATCTCGTATGCCGTCTTCTGCTTG
35	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCGTCTAATCTCGTATGCCGTCTTCTGCTTG
36	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTATCAGCAATCTCGTATGCCGTCTTCTGCTTG
37	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGGATGAATCTCGTATGCCGTCTTCTGCTTG
38	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGTTCTAATCTCGTATGCCGTCTTCTGCTTG
39	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCTGTCAATCTCGTATGCCGTCTTCTGCTTG
40	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCGTAGAATCTCGTATGCCGTCTTCTGCTTG
41	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTACGCAAATCTCGTATGCCGTCTTCTGCTTG
42	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGTGCGAAATCTCGTATGCCGTCTTCTGCTTG
43	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGAGAACAATCTCGTATGCCGTCTTCTGCTTG
44	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCTCGGTAATCTCGTATGCCGTCTTCTGCTTG
45	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCTAACGAATCTCGTATGCCGTCTTCTGCTTG
46	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCGAGTAAATCTCGTATGCCGTCTTCTGCTTG
47	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCACATAATCTCGTATGCCGTCTTCTGCTTG
48	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATAGACAATCTCGTATGCCGTCTTCTGCTTG
49	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGCTGAAATCTCGTATGCCGTCTTCTGCTTG
50	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAAGACTAATCTCGTATGCCGTCTTCTGCTTG
51	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAATCTGAATCTCGTATGCCGTCTTCTGCTTG
52	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTGGCATAATCTCGTATGCCGTCTTCTGCTTG
53	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTGAGCCAATCTCGTATGCCGTCTTCTGCTTG
54	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTCAATGAATCTCGTATGCCGTCTTCTGCTTG
55	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCATACAATCTCGTATGCCGTCTTCTGCTTG
56	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGACTGGAATCTCGTATGCCGTCTTCTGCTTG
57	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGAACTTAATCTCGTATGCCGTCTTCTGCTTG
58	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCTCCTGAATCTCGTATGCCGTCTTCTGCTTG
59	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTGAGAATCTCGTATGCCGTCTTCTGCTTG
60	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGAAGTAATCTCGTATGCCGTCTTCTGCTTG
61	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCAGTTCAATCTCGTATGCCGTCTTCTGCTTG
62	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCATACCAAATCTCGTATGCCGTCTTCTGCTTG
63	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGCGTTAATCTCGTATGCCGTCTTCTGCTTG
64	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACTTCGAATCTCGTATGCCGTCTTCTGCTTG
65	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAATGGAAATCTCGTATGCCGTCTTCTGCTTG
66	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGACTAGTAATCTCGTATGCCGTCTTCTGCTTG
67	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACCACAATCTCGTATGCCGTCTTCTGCTTG
68	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTGAGGAATCTCGTATGCCGTCTTCTGCTTG
69	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCCTGTAATCTCGTATGCCGTCTTCTGCTTG
70	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCACTAATCTCGTATGCCGTCTTCTGCTTG
71	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCAGGAAATCTCGTATGCCGTCTTCTGCTTG
72	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGATCGCAATCTCGTATGCCGTCTTCTGCTTG
73	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGAGTCAAATCTCGTATGCCGTCTTCTGCTTG
74	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTATGCAATCTCGTATGCCGTCTTCTGCTTG
75	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACGTATCAATCTCGTATGCCGTCTTCTGCTTG
76	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACGCTCGAATCTCGTATGCCGTCTTCTGCTTG
77	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCTCCAAATCTCGTATGCCGTCTTCTGCTTG
78	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGCAGAATCTCGTATGCCGTCTTCTGCTTG
79	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACACAGAAATCTCGTATGCCGTCTTCTGCTTG
80	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAGGTACAATCTCGTATGCCGTCTTCTGCTTG
81	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAGACGGAATCTCGTATGCCGTCTTCTGCTTG
82	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACGCTTAATCTCGTATGCCGTCTTCTGCTTG
83	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCGAGAATCTCGTATGCCGTCTTCTGCTTG
84	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACAACCAATCTCGTATGCCGTCTTCTGCTTG
85	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTACAAGATCTCGTATGCCGTCTTCTGCTTG
86	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTGTAGCCATCTCGTATGCCGTCTTCTGCTTG
87	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAAGCTAATCTCGTATGCCGTCTTCTGCTTG
88	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTGATCATCTCGTATGCCGTCTTCTGCTTG
89	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCATCAAGTATCTCGTATGCCGTCTTCTGCTTG
90	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCTGATCTCGTATGCCGTCTTCTGCTTG
91	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACATTGGCATCTCGTATGCCGTCTTCTGCTTG
92	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCACTGTATCTCGTATGCCGTCTTCTGCTTG
93	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTGGTCAATCTCGTATGCCGTCTTCTGCTTG
94	/5phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGCCTAAATCTCGTATGCCGTCTTCTGCTTG
95	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAAACATCGATCTCGTATGCCGTCTTCTGCTTG
96	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAACGTGATATCTCGTATGCCGTCTTCTGCTTG

Table 3 - Sequence of Barcoded P7 Adaptors

2.7 Next Generation Sequencing

For all whole genome, exome or targeted-capture sequencing, samples are sequenced on a 100 pair-ended flowcell on Illumina Hi-Seq 2000. For copy number profiling, samples are sequenced on 76 single-read flowcell. For the single cell sequencing in Chapter 6, cells are sequencing on Illumina Hi-Seq 3000.

2.8 Data Alignment and Processing

The FASTQ file containing all of the NGS data is demultiplexed into individual FASTQ files using our in-house software (deplexer.pl). Individual FASTQ files are aligned to the human genome reference assembly (HG18 for Chapter 3, HG19 for Chapter 4-6) using Bowtie 2, and they are converted to BAM files using SAMtools. BAM files are then processed by Picard to remove PCR duplicates. Re-alignment is performed around indel regions using the Genome Analysis Toolkit (GATK). Sequencing reads with mapping quality lower than 40 are removed. To calculate coverage metrics, we use an in-house Perl script (cal-coverage_metrics.pl), which uses BEDTools to get coverage depth at each site and to calculate overall coverage depth and coverage breadth. We use GATK to generate a multi-cell VCF file. We also use GATK to recalibrate variant quality scores. We filter mutations that are only detected in one single cell, as well as mutations in clustered regions, in which multiple mutations are detected within a 10-bp window. We annotate and classify the variants using ANNOVAR. The outline of the data processing pipeline, variant detection and annotation steps are shown in Figure 3.

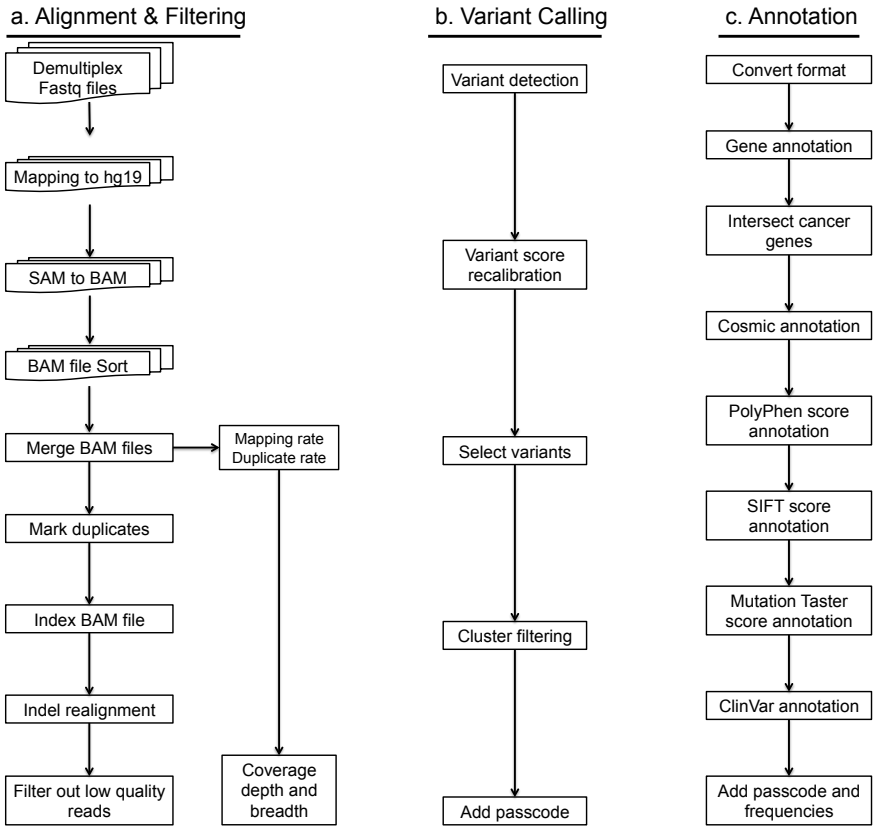

Figure 3 - Data Processing Pipeline

a. Alignment of sequencing reads to the reference genome and filtering by quality metrics. b. Detection of DNA variants and filtering of technical artifacts. c. Annotation of variants using integrated databases and protein damage-prediction algorithms.

(Modified and reproduced from Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7. Permission is not required, since Nature journal policy states “author retains the copyright to the published materials”,)

2.9 Calculation of Data Quality and Metrics

Coverage breadth is defined as the percent of genome (or targeted region) with at least one or more read. Coverage depth is defined as the average number of reads that each base of the genome (or targeted region) has.

The allelic dropout rate (ADR) is defined as the mean fraction of homozygous sites in the single cell samples (Hom_s) where the matched population reference sample is heterozygous (Het_p) at the same nucleotide site.

$$\boldsymbol{ADR} = \frac{1}{n}\sum_{i=1}^{n}\frac{Hom_s}{Het_p}$$

The false positive rate (FPR) is defined as the number of heterozygous sites in the single cell sample (Het_s) divided by the number of sites in the population reference sample that are homozygous (Hom_p) for the reference allele at the same nucleotide site.

$$\boldsymbol{FPR} = \frac{1}{n}\sum_{i=1}^{n}\frac{Het_s}{Hom_p}$$

Detection efficiency is defined as each variant as being detected if the reference allele is AB and the single cell data is either AB or BB. Detection efficiency can be calculated from the VCF4 variant files after the filtering steps are performed.

CHAPTER THREE – NUC-SEQ - SINGLE CELL WHOLE GENOME SEQUENCING

Chapter 3 – NUC-Seq – Single Cell Whole Genome Sequencing

Content of this chapter is based on: Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

(Marco L. Leung developed the single cell sequencing method and performed experiments for the breast cancer tumors in this study.)

3.1 Introductions and Rationale

In 2011, the first single cell DNA sequencing method, called Single Nucleus Sequencing (SNS), was published, demonstrating copy-number profiling of single tumor cell from breast cancer patients.[72] By using DOP-PCR, single cells are amplified and sparse-sequenced. By counting the sequencing reads throughout the genome, copy number profiling can be inferred, thus revealing tumor evolution by building phylogenetic tree.[72]

Although SNS is adequate for copy number detection, it is difficult to detect single-nucleotide variants (SNVs) using this method. It was shown that, even by performing deep sequencing, coverage breadth could not exceed more than 10%.[81] A new single cell amplification method is needed, in order to detect genome-wide mutations at single-nucleotide resolution.

To address this problem, we developed a novel method, called NUC-Seq, which can sequence the SNVs and indels of single cells. In this chapter, we demonstrate that, by using Φ29 polymerase, we can increase the genome coverage breadth up to 90%.[81] We also exploit the natural cell cycle, in which we select cells that are at the G2/M phase. This approach provides double genome content from 6 to 12 picograms, thus increasing the chance of polymerase amplifying both alleles. We limit the MDA reaction time to 80 minutes. This prevents the over-amplification of false positive errors created by the polymerase. Lastly, we applied NUC-Seq to sequence single cells from two breast cancer patients to understand intratumor heterogeneity and tumor evolution.

3.2 Results

3.2.1 Whole-Genome Sequencing Using G2/M Nuclei

We developed the single cell whole-genome sequencing method by combining flow-sorting of nuclei, multiple displacement amplification, quality control, library construction by tagmentation or dA tailing cloning and next generation sequencing. (Figure 4)

First, nuclear suspensions were prepared and stained by mixing cells with DAPI-NST buffer. Using a flow-sorter, we can determine the ploidy distribution of the nuclear suspension. The G2/M peak is gated and single nuclei were deposited into each well of a 96-well plate. Cells are then lysed and incubated with Φ29 polymerase to perform multiple-displacement-amplification for a limited isothermal time frame. A quality control step is performed on amplified DNA to

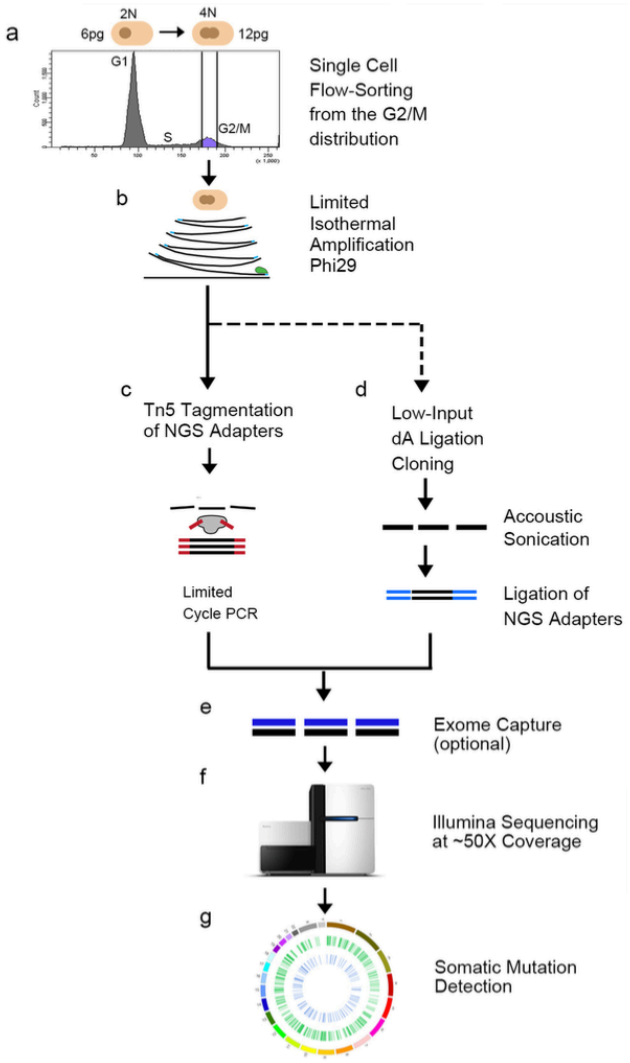

Figure 4 - NUC-Seq Method Overview

a. Nuclear suspensions were prepared and stained with DAPI for flow-sorting, showing distributions of ploidy. The G2/M distribution was gated and single nuclei were deposited into wells. b. Cells were lysed and incubated with the Φ29 polymerase to perform multiple-displacement-amplification for a limited isothermal time-frame. c.d. Sequence libraries were prepared using one of two methods: Tn5 tagmentation (c), or low-input TA ligation cloning (d). e. Exome capture was optionally performed to isolate gDNA in exonic regions. f. Libraries

were sequenced on the Illumina HiSeq 2000 system. g. Somatic mutations were detected using a custom processing pipeline.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

ensure even amplification and human origin. (Figure 5) We found that, in G1/0 single cells, only 25.58% (11/43) of the cells show full amplification of the chromosomes, whereas G2/M cells have 45.34% (39/86). Sequence libraries are prepared on cells that pass quality control using one of two methods: Tn5 tagmentation or low-input TA ligation cloning. (see Chapter 2 - Methods and Materials) Libraries are then sequenced on the Illumina HiSeq 2000 system. Somatic mutations are called using processing pipeline. (See Chapter 2 - Methods and Materials)

3.2.2 Method Validation in a Monoclonal Cancer Cell Line

We first validate our NUC-Seq method using a breast cancer cell line (SK-BR-3). In our previous study, we have shown that SK-BR-3 was a genetically monoclonal cell line.[72] We performed Single Nucleus Sequencing on 50 single SK-BR-3 cells and calculate the copy number profiles at 220-kilobase resolution. (Figure 6) We found that the major copy number aberrations (amplifications of *MET*, *MYC*, *ERBB2*, *BCAS1*, and deletion in *DCC*) were stable across all 50 cells. We also performed deep-sequencing the whole genome of SK-BR-3 cell population at high coverage depth (51x) and breadth (90.40%) and detected single-nucleotide variants (SNVs), copy number aberrations (CNAs) and structural variants (SVs) using our processing pipeline. We filtered the variants using dbSNP135 and identified non-synonymous SNVs and SVs. (Figure 7)

We then applied NUC-Seq to sequence the whole genomes of two single SK-BR-3 cells (named SK-1 and SK-2). Specifically, we sequenced single cells that are in G2/M phase of the cell cycle. We calculated the coverage depth and

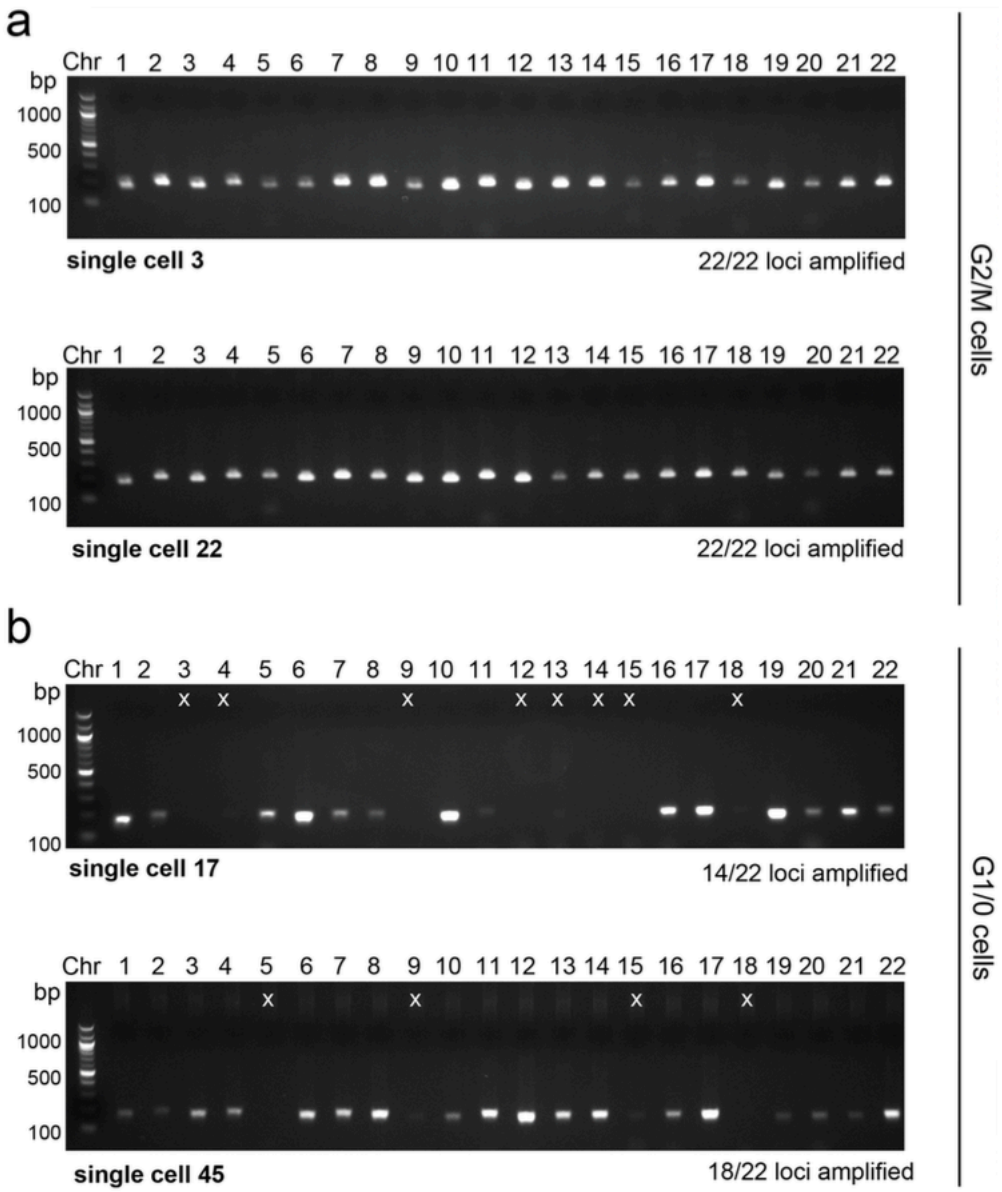

Figure 5 - Evaluation of WGA Efficiency Using Chromosome-Specific Primers

Whole genome amplified DNA from each single cell was used to perform PCR quality control experiments to determine WGA efficiency. For each cell, 22 reactions were performed using primer pairs that target each autosome and

resulting 200bp PCR product were separated by gel electrophoresis. Two single nuclei from G2/M gate (a) and G1/0 gate (b) are shown. ‘X’ on the gel represents negative PCR reactions.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that “author retains the copyright to the published materials.”)

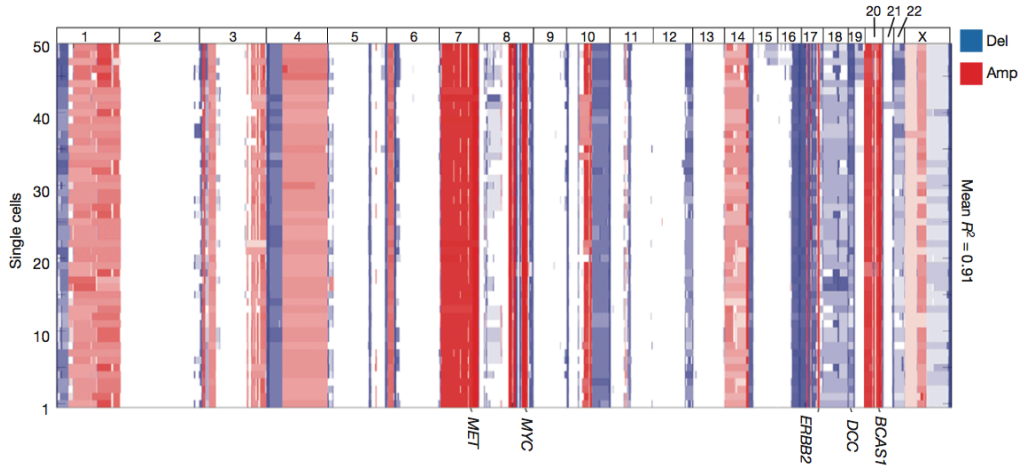

Figure 6 - Copy Number Heatmap of 50 Single SK-BR-3

Each row represents a single SK-BR-3 cell. Chromosome 1-Y are organized from left to right. Blue represents copy number loss and red represents copy number gain.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

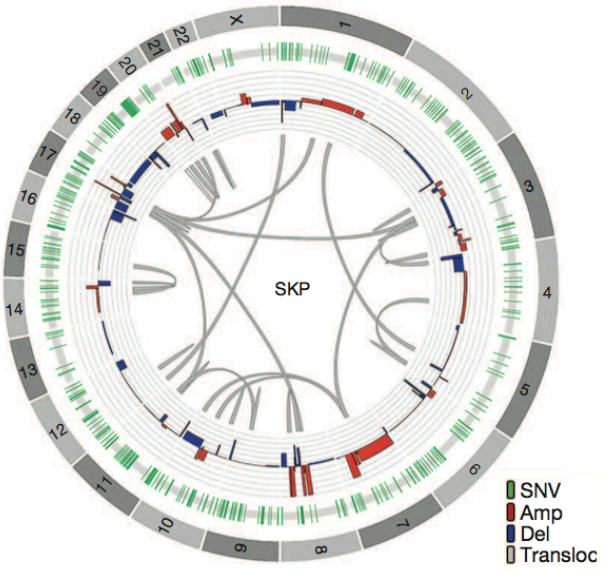

Figure 7 - Bulk Sequencing of SK-BR-3 Cell Line

The Circos plot represent the sequencing result of SK-BR-3 bulk whole genome sequencing. The outer circle shows the chromosome location. Next inner circle shows the SNVs and indels. The next inner circle shows the copy number aberrations. The most inner circle shows large scale chromosomal translocations.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600. Copyright permission is not required since Nature journal policy states that “author retains the copyright to the published materials.”)

coverage breadth, and compared to the population (SKP). We found that both SK-BR-3 cells achieved high coverage depth (61x) and high coverage breadth (83.7%). (Figure 8, 9)

Next, we calculated the allelic dropout rate (ADR) and false positive rate (FPR) by comparing single cell variants to the population data. Our analysis suggests that NUC-Seq generates low allelic dropout rate (9.73%) compared to previous studies (7-46%).[75] We also achieved low false positive error rates for point mutations (1.24×10^{-6}), equivalent to 1-2 errors per million bases, which represents a major technical improvement over previous methods (FPR = 2.52×10^{-5} and 4×10^{-5}).[75,79]

3.2.3 Single Cell Sequencing of Breast Tumors

We then selected tumors from two breast cancer patients for population and single cell sequencing. We first investigated an invasive ductal carcinoma from an estrogen-receptor positive ($ER^+/PR^+/Her2^-$) breast cancer patient. We flow-sorted millions of nuclei from the aneuploid G2/M peak (6N) and from matched normal tissue for population sequencing. (Figure 10) We also flow-sorted 50 single nuclei for copy number profiling, 4 nuclei for whole-genome sequencing and 59 nuclei for exome sequencing. After filtering germline variants, we identified a total of 4,162 somatic SNVs in the aneuploid tumor cell population. Among these SNVs, we identified 12 nonsynonymous mutations, which we validated by exome sequencing. Several non-synonymous mutations

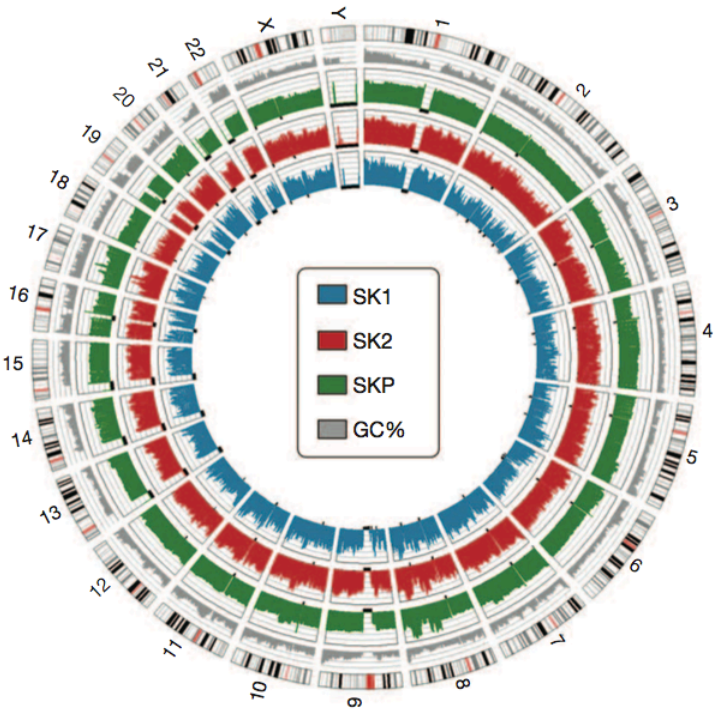

Figure 8 - Coverage Depth for Bulk and Single Cell Sequencing

The three inner circles represent the coverage depth of two single cells (blue and red) and population sequencing (green). The coverage performance is corresponded to the GC content of the genome (grey)

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

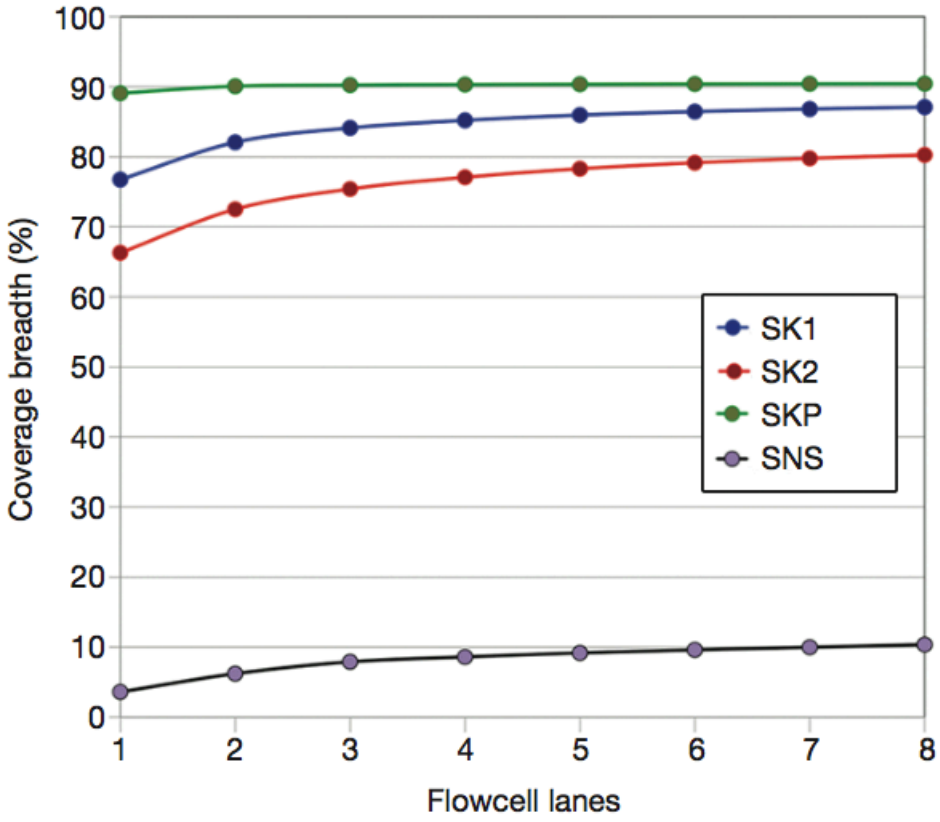

Figure 9 - Coverage Breadth for Bulk and Single Cell Sequencing

This line graph shows that the coverage breadth performance of single cells is comparable to the performance of population. This SCS method shows a large improvement of our previous method, single nucleus sequencing.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600. Copyright permission is not required since Nature journal policy states that “author retains the copyright to the published materials.”)

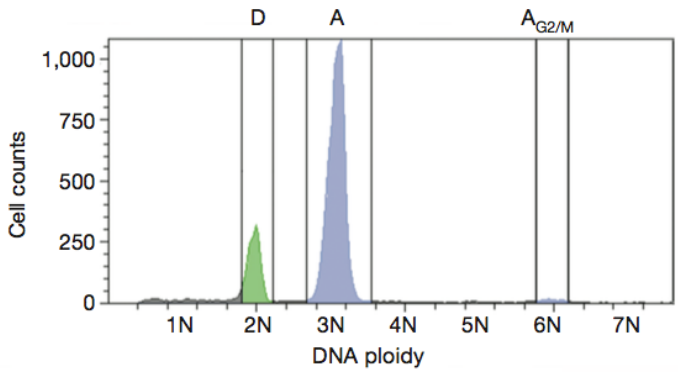

Figure 10 - Ploidy Distribution of an Estrogen-Receptor Positive Breast Tumor

Single cells were sorted from the G2/M aneuploid peak, whereas populations are sort from the diploid and aneuploid peak.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

occurred in cancer genes, including *PIK3CA*, *CASP3*, *FBN2* and *PPP2R5E*. (Figure 11).

To investigate copy number diversity, we performed single nucleus sequencing on 50 single nuclei. We constructed a neighbor-joining tree, which showed that single tumor cells shared highly similar CNAs, representing a monoclonal population. (Figure 12) Next, we performed whole-genome sequencing of four single tumor nuclei at high coverage breadth and depth. From this data, we identified three classes of mutations: (1) clonal mutations, detected in the population sample and in the majority of single tumor cells; (2) subclonal mutations, detected in two or more single cells, but not in the bulk tumor; and (3) *de novo* mutations, found in only one tumor cell. The *de novo* mutations are difficult to distinguish from technical errors and were therefore exclude from our initial analysis. In total, we detected 12 clonal non-synonymous mutations and 32 subclonal mutations. (Figure 13) Many subclonal mutations occurred in intergenic regions; however, two mutations (*MARCH11* and *CABP2*) were found in coding regions.

To identify additional subclonal mutations, we performed single nuclei exome sequencing on a larger set of cells (47 tumor cells and 12 normal cells). Each nucleus was sequencing at 46.78x coverage depth and 92.77% exome coverage breadth, from which somatic mutations were detected. The mutations were clustered and sorted by frequency to construct a heatmap. (Figure 14) As expected, the 17 clonal mutations identified by population sequencing were

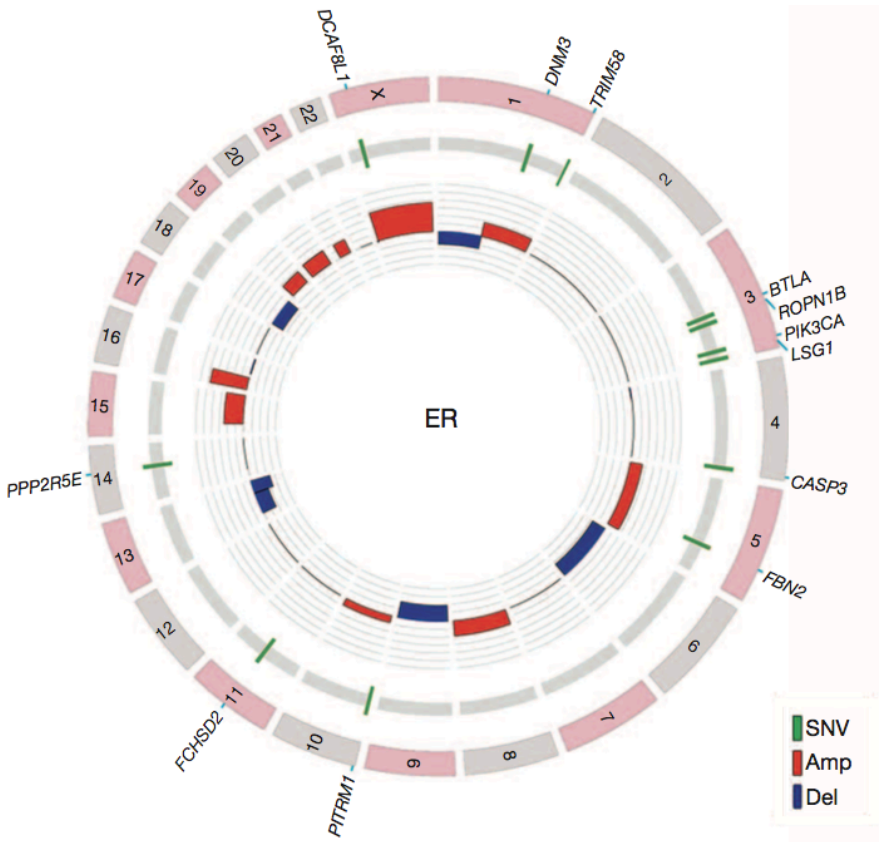

Figure 11 - Circos Plot of Mutations and CNAs in ER+ Breast Tumor

SNVs and CNVs are detected in the bulk sequencing of the tumor.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

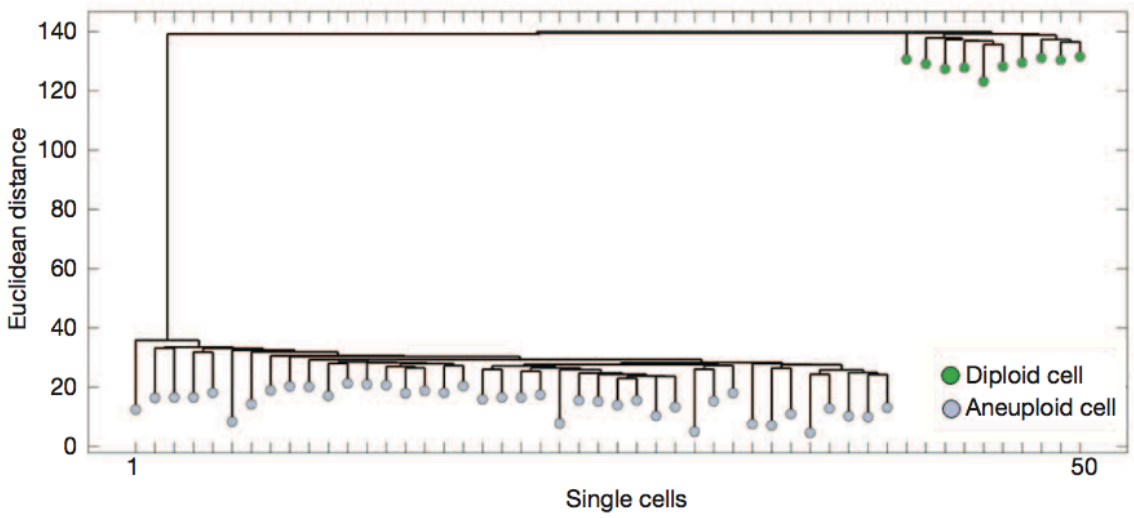

Figure 12 - Neighbor-Joining Tree of Single Cell Copy Number Profiles

Neighbor-joining tree of integer copy number profiles from single diploid and aneuploid cells, rooted by the diploid node.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

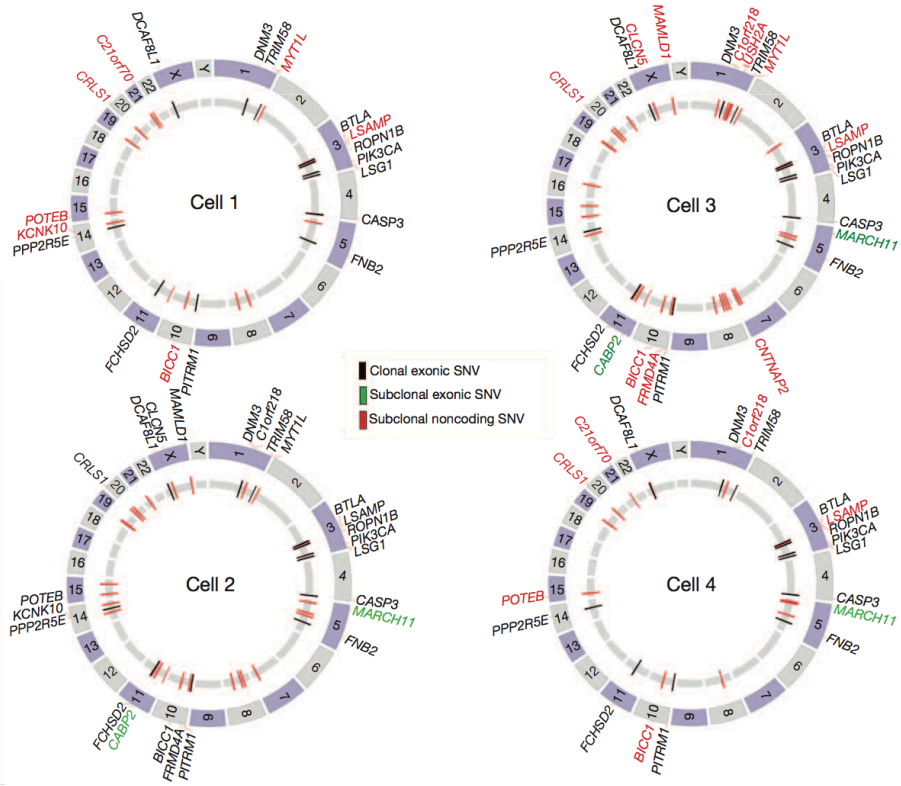

Figure 13 - Circos Plots of Single Cell Whole Genome Profiles

These circus plots of whole-genome single cell sequencing data showing mutations detected in two or more cells. Black, green and red represent clonal exonic, subclonal exonic, and subclonal noncoding SNVs, respectively.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

present in many of the single tumor cells. However, we also identified 22 new subclonal mutations. In contrast, only a single subclonal mutation was detected in the 12 normal cells. (Figure 14)

Next, we proceeded to analyze a triple-negative ($ER^-/PR^-/Her2^-$) breast cancer (TNBC). From population sequencing of the bulk tumor and matched normal tissue, we identified 374 non-synonymous mutations, which is significantly higher than the ER tumor. A number of mutations occurred in cancer genes, including *PTEN*, *TBX3*, *NOTCH2*, *JAK1*, *ARAF*, *NOTCH3*, *MAP3K4*, *NTRK1*, *AFF4*, *CDH6, SETBP1*, *AKAP9*, *MAP2K7*, *ECM2* and *ECM1*. (Figure 15) Many of these mutations were previously reported in the TCGA breast cancer cohort.[82]

To investigate genomic diversity at single cell resolution, we performed copy number profiling and exome sequencing. We flow-sorted 50 single nuclei from the hypodiploid, diploid and aneuploid ploidy distributions for copy number profiling using SNS. Neighbor-joining revealed two distinct subpopulations of tumor cells (A and H) in addition to the normal diploid cells. (Figure 16) The single cell copy number profiles were analyzed using clustered heatmaps, which showed highly similar rearrangements within each subpopulation, but were distinguished by two large deletions on chromosome 9 and 15.

We then flow-sorted 16 single tumor nuclei from the G2/M peaks of hypodiploid and aneuploid, as well as 16 nuclei for exome sequencing. Non-synonymous point mutations were used to perform hierarchical clustering and multi-dimensional scaling (MDS). As expected, the 374 clonal non-synonymous

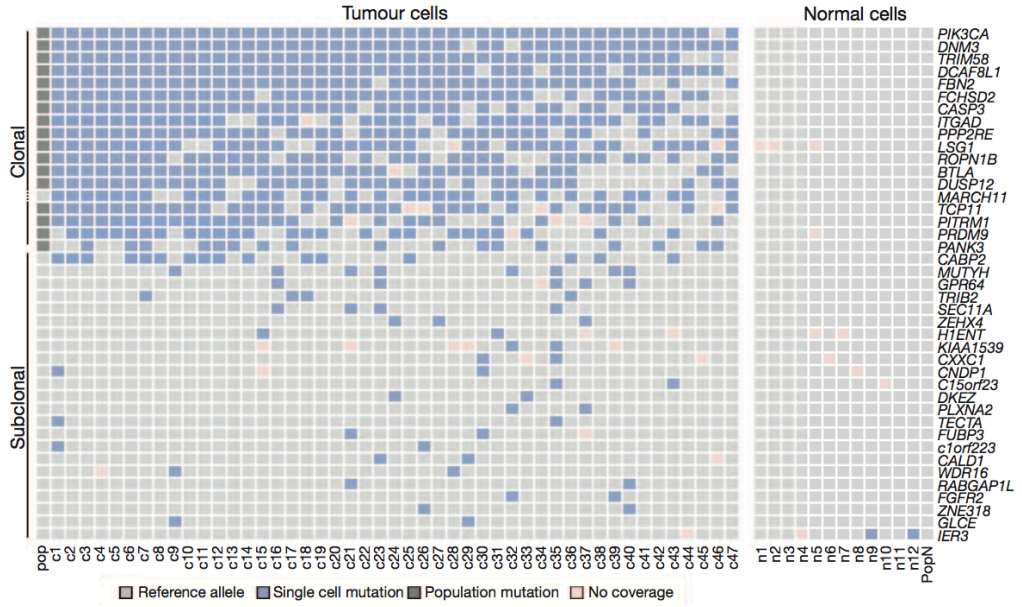

Figure 14 - Mutations Detected in Single Cells Exome Sequencing

Heatmap of coding mutations detected by single-nuclei exome sequencing. Mutations detected in whole-genome sequencing (pop) and exome sequencing (c) are displayed.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that “author retains the copyright to the published materials.”)

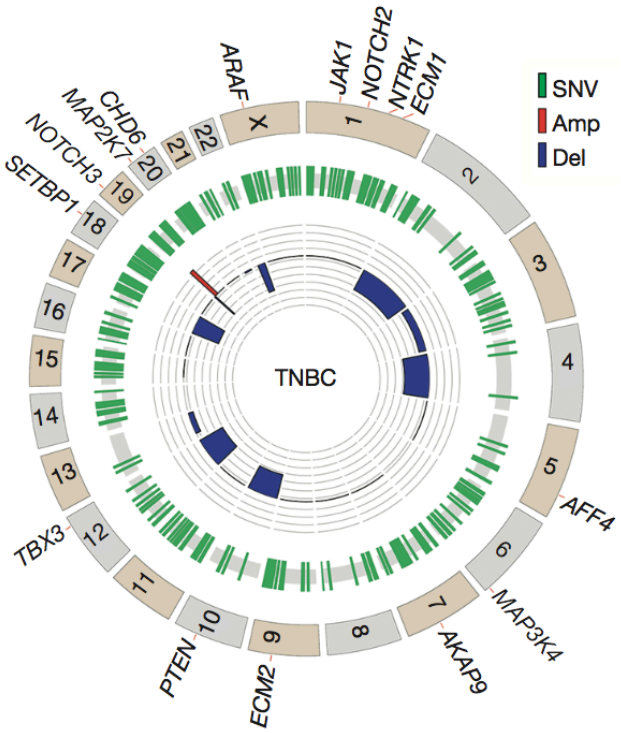

Figure 15 - TNBC Mutations Detected in Population Sequencing

Circos plots of mutations and CNAs detection by population sequencing of the triple negative breast cancer, with cancer genes on the outer ring.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

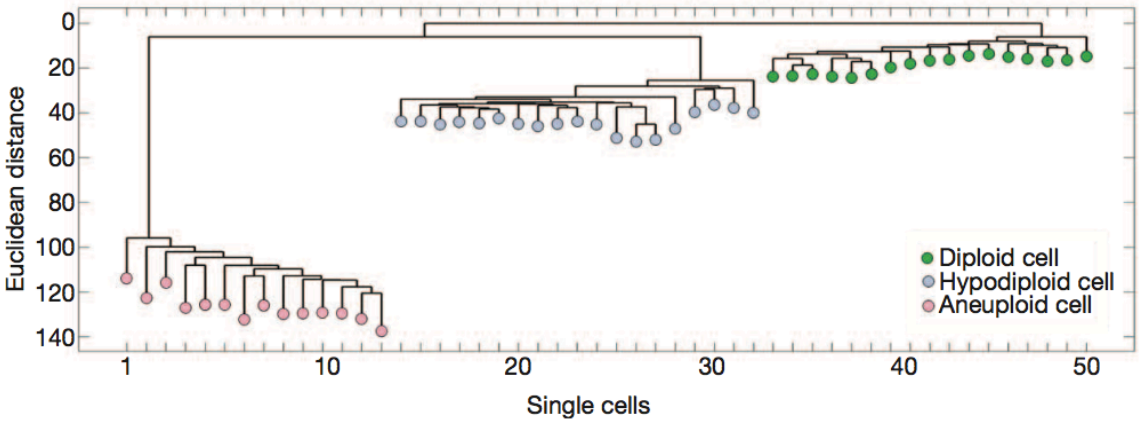

Figure 16 - Neighbor-Joining Tree of TNBC Single Cell Copy Number Profiles

Neighbor-joining tree of 50 single cell integer copy number profiles, rooted by the diploid node. There are three distinct subpopulations, clustered by their own ploidy profiles.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

mutations detected by bulk sequencing were found in the majority of the single tumor cells, however, we also identified 145 additional subclonal non-synonymous mutations that were not detected in the bulk tumor. MDS identified 4 distinct clusters, corresponding to three tumor subpopulations and the normal cells. (Figure 17) Hierarchical clustering showed that many of the subclonal mutations occurred exclusively in one subpopulation. (Figure 18) Many of the subclonal mutations were predicted to have damage protein function by both POLYPHEN and SIFT.[83,84]

3.3 Discussion

In this chapter, we have presented a novel single cell whole genome sequencing method, called NUC-Seq. By combining flow-sorting, multiple-displacement-amplification, quality control, library preparation, next-generation sequencing, we are able to generate whole-genome data from single cells with high coverage breadth and high coverage depth. We are able to achieve high coverage depth (61x) and breadth (83.0%). Compared to our previous method, which uses sparse sequencing to detect CNVs, this improvement in coverage is particularly important because it detects mutations in single-nucleotide resolution. We demonstrate this technique to reveal subclonal mutations and intratumor heterogeneity in two breast tumors. We are able to detect rare mutations that would not be otherwise detected using conventional population sequencing techniques.

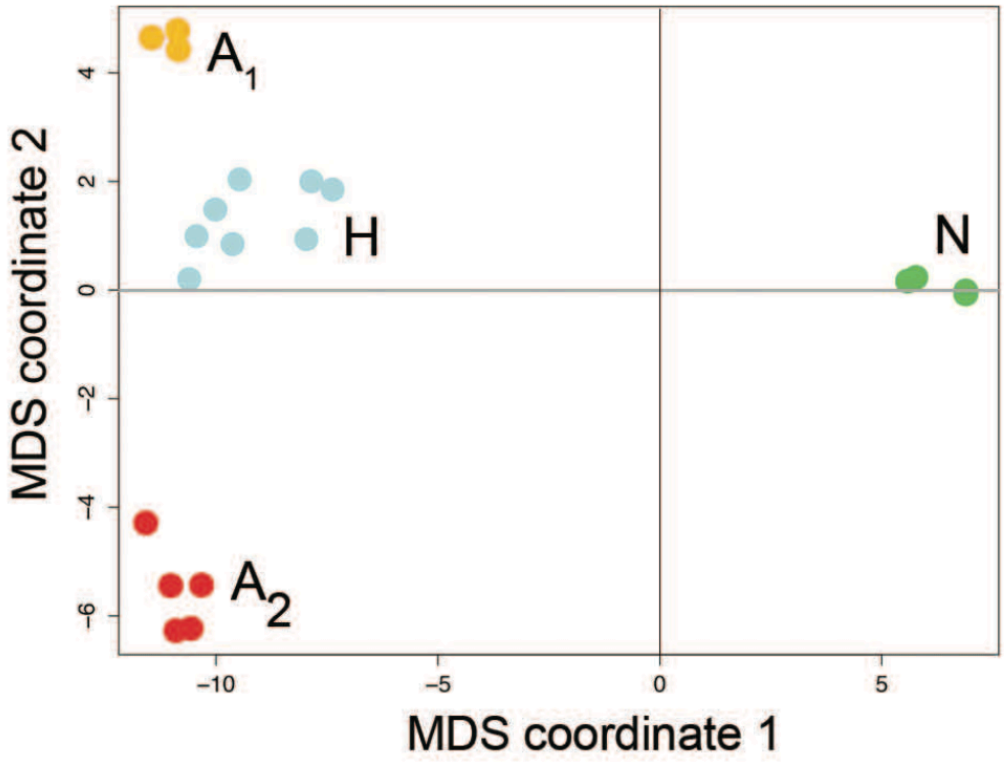

Figure 17 - TNBC Multi-Dimensional Scaling (MDS) Plot

Multi-dimensional scaling plot of the nonsynonymous mutations from the single-nuclei exome sequencing data in the TNBC.

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

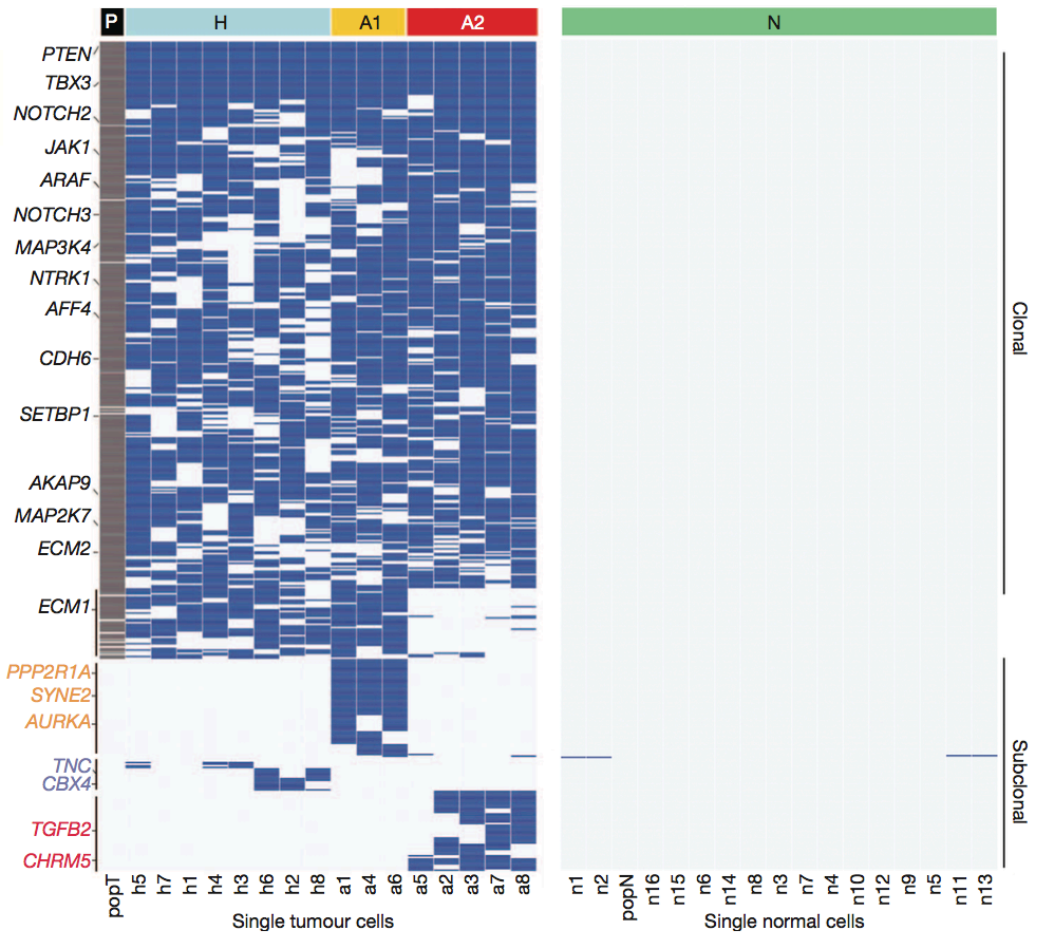

Figure 18 - Mutations Detected in TNBC Single Cell Sequencing

Clustered heatmap of the nonsynonymous point mutations detected by single nuclei exome sequencing and populations

(Modified and reproduced from Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam & Nicholas E. Navin. *Nature*. 2014. 512(13500:155-160). doi:10.1038/nature13600

Copyright permission is not required since Nature journal policy states that "author retains the copyright to the published materials.")

CHAPTER FOUR – SNES – SINGLE NUCLEUS EXOME SEQUENCING

Chapter 4 – SNES – Single Nucleus Exome Sequencing

Content of this chapter is based on: Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

4.1 Introductions and Rationale

While single cell sequencing methods provide an alternative approach in detecting rare mutations by sequencing single cells, further technical improvements are still needed to decrease the error rates. Following the publication of our SNS method, two recent methods were developed that use multiple displacement amplification (MDA) and multiple-annealing-looping-based-amplification-cycles (MALBAC) to increase the coverage breadth during WGA.[75,79] While pioneering, these studies increased coverage breadth at the cost of introducing high false positive and false negative error rates, due to excessive over-amplification of the DNA from a single cell from 6 picograms to microgram concentrations. Consequently, it was necessary to call variants across most of the single cells to reduce the high false positive (FP) technical errors, which is equivalent to sequencing the bulk tissue *en masse*.

To mitigate technical errors, we developed NUC-Seq, which utilizes G2/M cells to perform single-cell genome sequencing.[81] While this approach was suitable for analyzing highly proliferative cells, such as cancer cells, it was not suitable for analysis of normal cells or slowly dividing populations. To address this problem, we developed a new approach called single nucleus exome sequencing (SNES) that was built upon our previous method.

In this chapter, we describe the development of SNES and address the following problems in existing single cell genome sequencing methods.

1. Error rates are high due to over-amplification of single cell DNA.
2. NUC-Seq focuses only on single cells in G2/M phase.
3. It is expensive to sequence cells using commercial kits.

4.2 Results

4.2.1 Experimental Approach and Quality Control Assays

The experiment approach of SNES is illustrated in Figure 19. It is similar to NUC-Seq with minor modifications and improvements. After nuclear suspensions are prepared from cells using the DAPI-NST lysis buffer, flow-cytometry is used to sort cells in G1/0 or G2/M phase. After depositing single nuclei into individual well of a 96-well plate, we use phi29 polymerase to amplify single cell. To determine the optimal isothermal time-frame, we performed time-series MDA reactions using G1/0 and G2/M cells over 8 hours. (Figure 20) From this amplification curve, we determine that 120 minutes to be the minimum time-frame required to generate approximately 500 ng of DNA from a single cell,

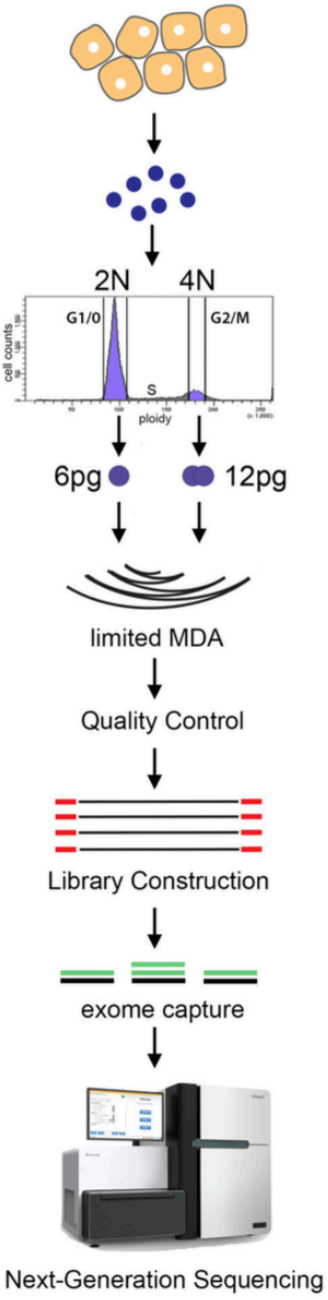

Figure 19 - SNES Experimental Procedure

Nuclear suspensions were prepared from tissues, stained with DAPI and flow-sorted. Single nuclei were isolated by gating the G1/0 or G2/M ploidy

distributions and deposited singly into a 96-well plate. Multple-displacement-amplification is performed using Φ29 to perform WGA.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

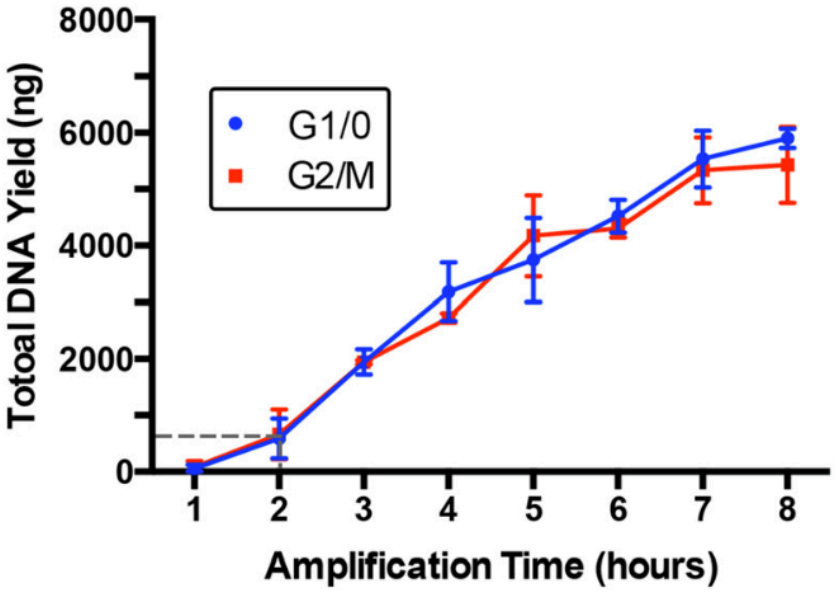

Figure 20 - Amplification Curve of Single Nuclei Using phi29 Polymerase

Time course of WGA showing total DNA yield of from single nuclei of both G1/0 or G2/M phase

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

providing sufficient input material for constructing libraries, exome capture, and performing the necessary quality control assays.

Prior to library construction, we evaluate WGA efficiency by performing qPCR on each single nucleus WGA reaction using a set of 22 primary pairs that target each chromosome independently. (See **Chapter 2**) Single nuclei with 22/22 amplicons were selected for subsequent library construction and next-generation sequencing. Our data showed that G2/M cells resulted in an improvement over G1/0 cells for WGA efficiency, with 70% (14/20) single cells having the full set of chromosomes amplified in G2/M cells compared to 15% (3/20) in G1/0 cells. (Figure 21)

Single cells that passed quality control for WGA were used to construct sequencing libraries using a low-input TA cloning protocol starting with 100ng of input material. During library construction, a unique 6-bp barcode was added to each single-cell libraries together into one reaction for exome capture and next-generation pair-end sequencing on the HiSeq2000 system using 100 paired-end cycles.

4.2.2 Measuring Coverage Performance and Uniformity

To determine the coverage performance and error rates of SNES, we used a normal isogenic female fibroblast cell line (SKN2), in which we assume that the variants present in a single cell will be highly similar to the reference population sample. Any deviations from the reference variants were considered to be technical errors, and were used to calculate the error rates. We sequenced the population of cells at high coverage depth (59x) and breadth (99.76%) to

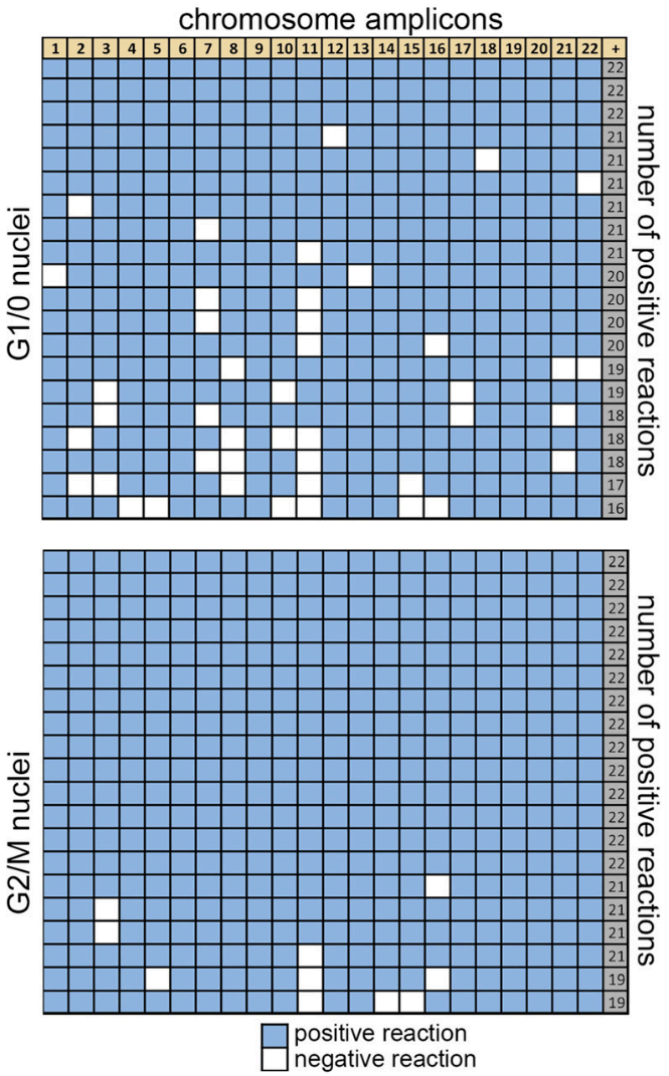

Figure 21 - qPCR Panel for Single Nuclei

Quality control assay using a panel of 22 chromosome-specific qPCR primers to determine the WGA amplification efficiency of each single nucleus

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

obtain a reference set of whole-genome variants. We then applied SNES to sequence nine single cells that were gated from the G1/0 stage of the cell cycles and 10 single cells from the G2/M stage. We aligned the single-cell data to the human genome using our processing pipeline and eliminated sequencing reads with multiple mappings and PCR duplicates. As expected, all of the single cells showed very similar coverage depth distributions, irrespective of whether they were gated from the G1/0 or G2/M distribution, which is important for the subsequent comparisons. (Figure 22)

In order to assess coverage performance, we calculated coverage breadth (sites with ≥ 1x coverage) (Figure 23) and coverage uniformity (evenness) (Figure 24). Our data suggest that coverage breadth (≥1x) significantly (p = 0.0021, t-test) increased in the G2/M cells (95.94%, ±0.005 sem) relative to the G1/0 cells (89.60%, ±0.018 sem) (Figure 23). This results in the number of sites with sufficient coverage depth for variant calling at 73.54% in G1/0 cells compared to 84.34% in G2/M cells. To assess coverage uniformity, we plotted the fraction of the exome covered as a function of coverage depth (Figure 25). These plots show that the G2/M cells achieved more even coverage uniformity at sites with low coverage depth compared to the G1/0 cells.

To further investigate coverage uniformity, we calculated Lorenz curves and plotted data for perfect uniformity, a genomic DNA population sample and mean data for the G1/0 and G2/M single cells, as well as data from our previous SNS method (Figure 24). These curves show a large improvement in coverage uniformity using G2/M cells compared to the G1/0 cells, and both showed vast

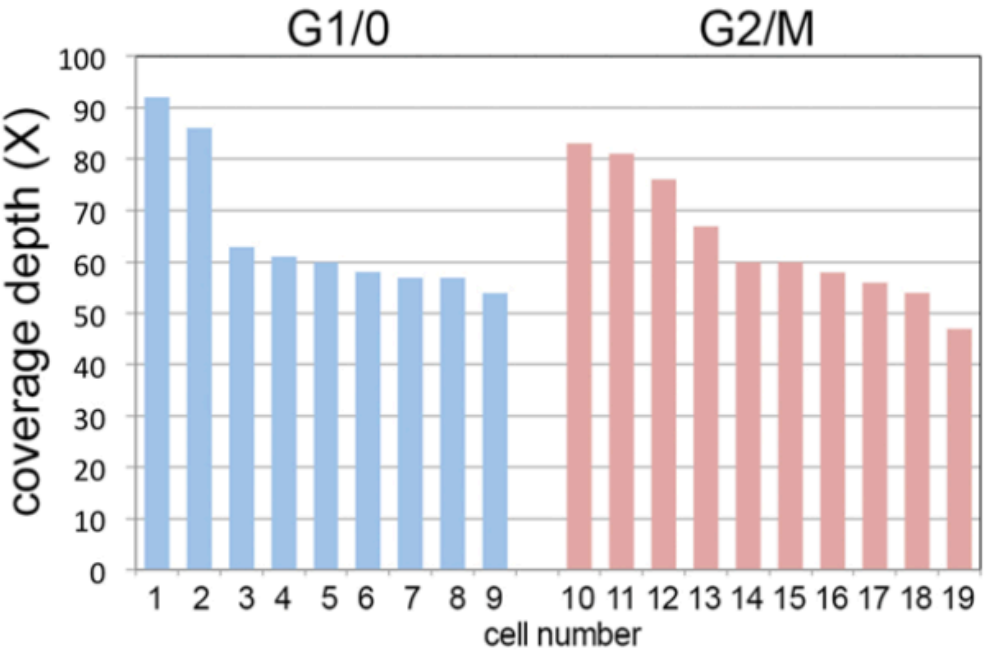

Figure 22 - Coverage Depth of Single Nuclei

9 nuclei from the G1/0 phase and 10 from the G2/M phase were sequenced. There is no significant difference in coverage depth between the two groups.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

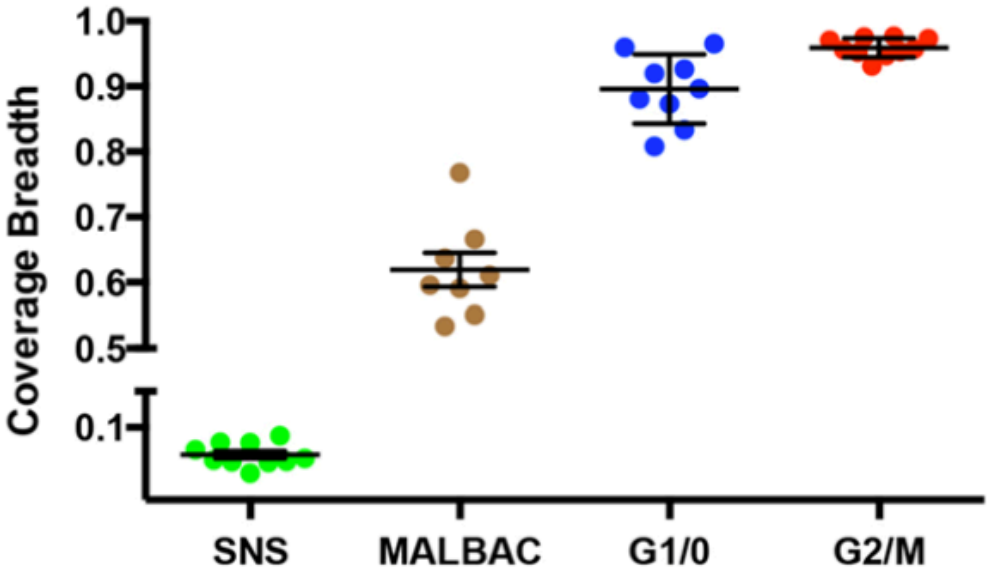

Figure 23 - Coverage Breadth of Single Nuclei

Coverage breadth data for exome region of G1/0 and G2/M single cells compared to previous studies using SNS and MALBAC. Error bars show SEM.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.
Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

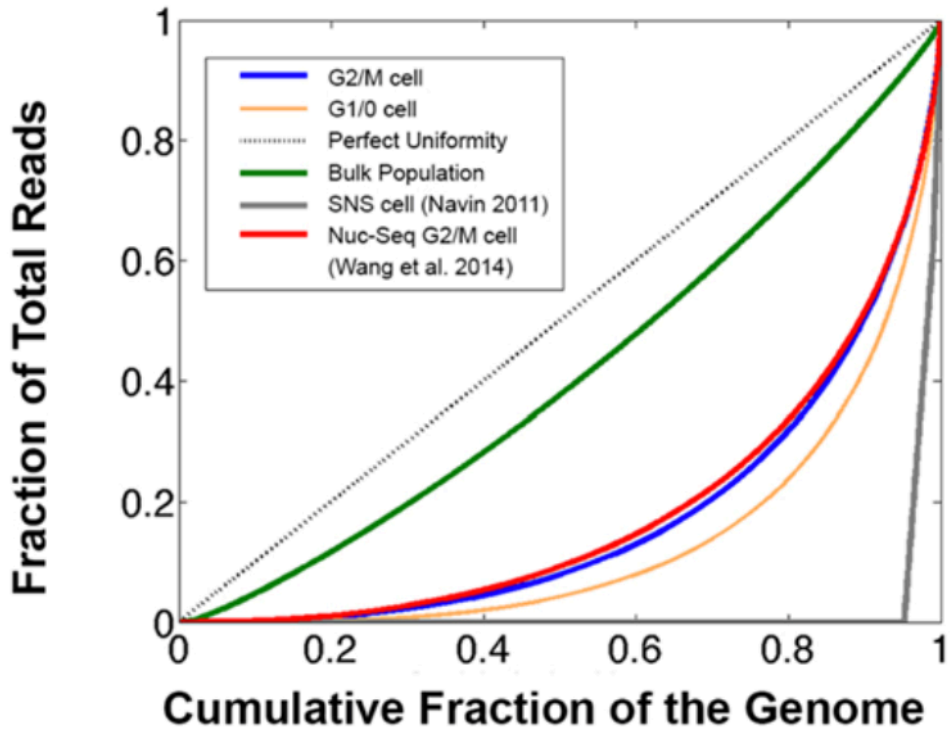

Figure 24 - Coverage Uniformity Comparison

Lorenz curves of coverage uniformity, showing values for perfect coverage, milliions of SKN2 reference cells, NUC-Seq single cell, single cells from G1/0 and G2/M distributions, and SNS cell.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

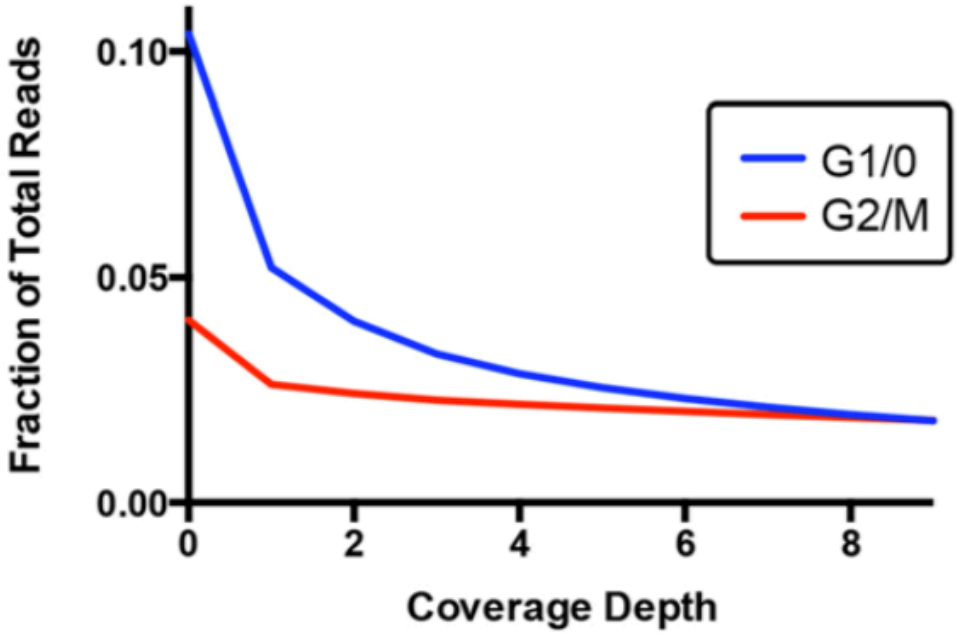

Figure 25 - Coverage Distribution for Sites with Low Coverage in G1/0 and G2/M Single Cells

This plot shows that there are more sequencing reads at the low coverage depth for G1/0 cells.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

improvements over our previous SNS approach. We also calculated the on-target performance for data in the exome region of single cells, and found very high percentages (mean = 67.33%) for G1/0 and G2/M cells, which are equivalent to previous reports (55% to 85%) of exome capture efficiencies using millions of cells.[85]

4.2.3 Estimating Technical Error Rates

To calculate the technical error rates, we filter the reads by mapping quality, base quality, and clustered regions. We then perform local realignment around indels. From these data, we identified SNVs and indels using the Unified Genotyper (GATK), following our processing pipeline. Major sources of technical errors that occur during WGA include the ADR and the FP error rates. (Figure 26) Previous studies have reported very high ADR (43.09%) in single-cell exome sequencing data.[76,86] In comparison, our data show that SNES significantly ($p = 7e-4$, *t*-test) reduced the ADR to 30.81% (±0.013, sem) in G1/0 cells and 21.52% (±0.019, sem) in G2/M cells (Figure 27). These calculations are based on sites in which both the single cells and population sample have sufficient (≥ 6x) coverage depth (in order to eliminate sites with low coverage in which WGA did not necessarily lead to allelic dropout). An alternative approach for calculating the ADO includes all heterogeneous sites in the population and single cell sites regardless of coverage depth, which results in an ADR of 43.84% for G1/0 cells, and 27.21% for G2/M cells.

Next, we calculated the FP error rate, which is caused by the infidelity of the phi29 polymerase (error rate = 1e-7) during isothermal amplification.[87] From

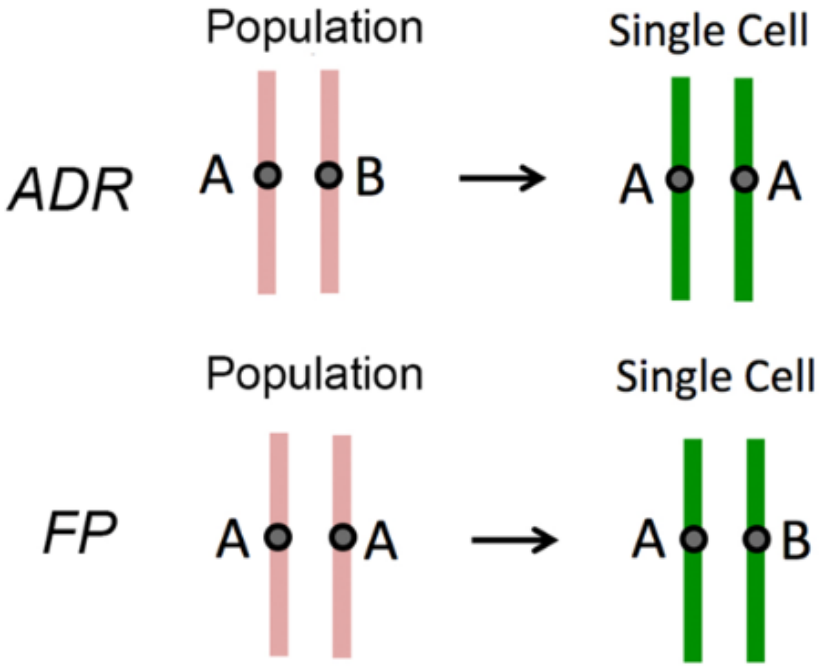

Figure 26 - Allelic Dropout Rate (ADR) and False Positive (FP)

Allelic dropout is defined as the loss of one allele when there are two alleles present in the population sequencing sample. False positive is defined as the detection of an allele that artificially created during amplification and not present in the population sequencing.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

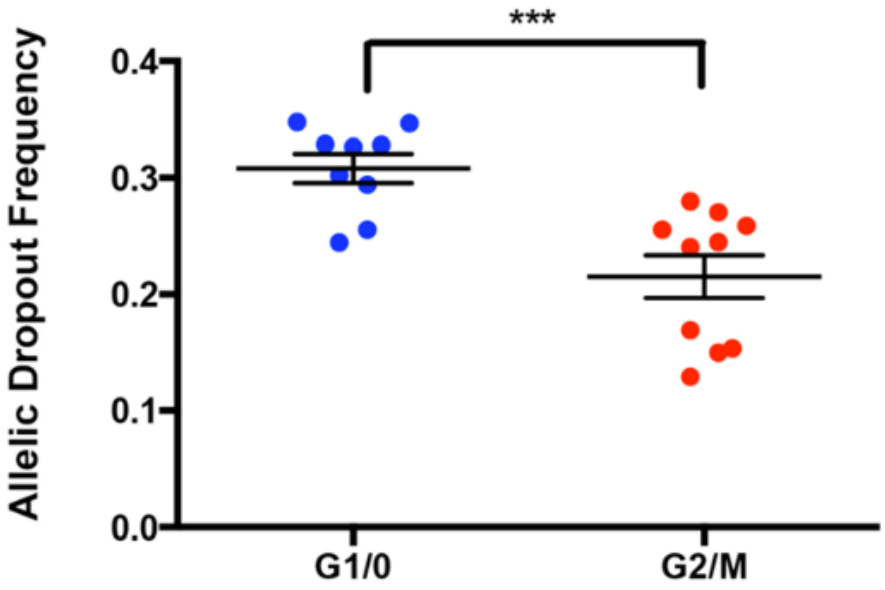

Figure 27 - Allelic Dropout Rate Comparing G1/0 and G2/M Cells

9 G1/0 and 10 G2/M cells are compared for their ADR. G1/0 cells have average of 30.81% ADR and G2/M cells have average of 21.52% ADR.

*** = P≤0.001

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

our data, we calculated a FP error rate of 3.2e-5 for SNVs, which is equivalent to 32 errors per megabase. This FP error rate is higher than our previous estimates for whole-genome single-cell sequencing with NUC-Seq, but can be explained by the increased isothermal WGA timeframe and additional PCR cycles required to generate sufficient DNA for exome capture and enrichment. We investigated the spectrum of the FP errors and found that 82.3% were C > T and G > A transitions, showing a significant bias relative to the normal transition and transversion spectrum in the population of fibroblast cells. (Figure 28) Importantly, we found that the majority of the FP errors occurred at random sites in the genomes of single cells, with few mutations occurring at recurrent sites in two or more cells. This distribution allows the FP erros rates to be mitigated by calling mutations in two (FP: $(3.2e\text{-}5)^2$ = 1.02e-9) or more (FP: $(3.2e\text{-}5)^n$) single cells. Using two or more cells in variant calling is possible in most single-cell studies, which normally seek to analyze large numbers of cells.

We also investigated the distribution of allelic dropout events in the single-cell data. By comparing the allelic dropout events from both alleles, our data showed that there is a slight bias towards AB -> BB dropout event, when compared to AB -> AA events in both the G1/0 and G2/M events. (Figure 29) We suspect that this bias is likely due to mismatch hybridization inefficiency of the exome capture probes to the B alleles, they were designed for the A allele sequence (reference human genome assembly). Next we examined the distribution and recurrence of allelic dropout events by examining their frequency across multiple single cells. Our data show that in contrast to the random

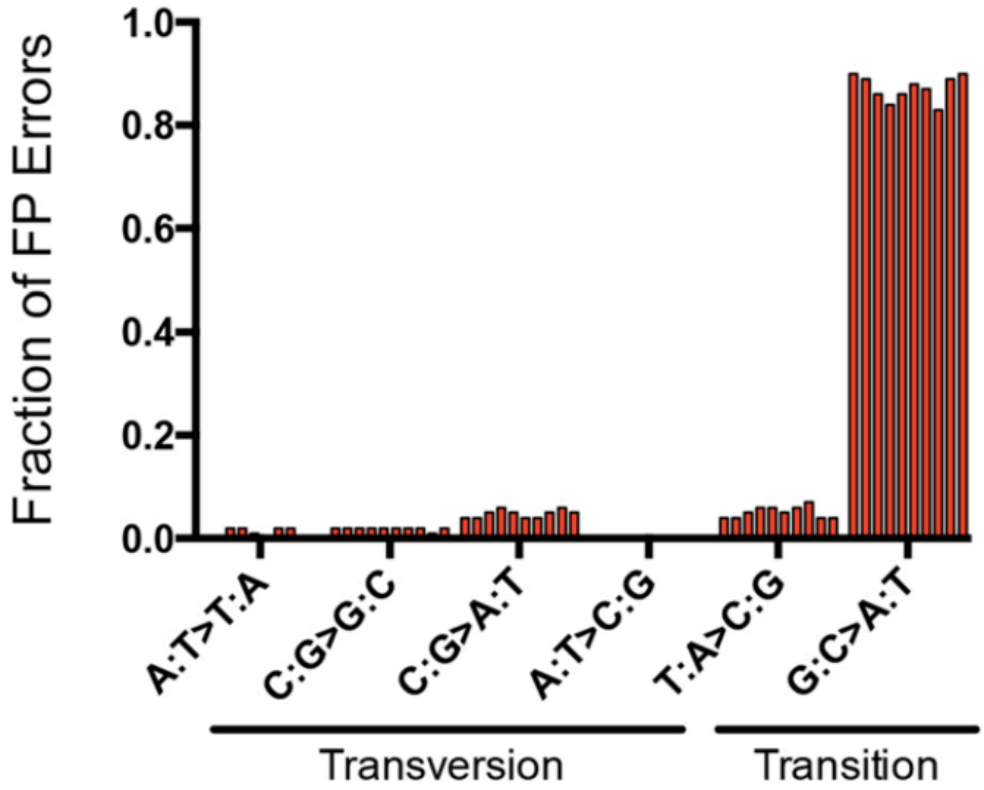

Figure 28 - Spectrum of Single Nucleotide Variants

82.3% of single nucleotide variants detected in the G2/M single cell data have G>A and C>T transition.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

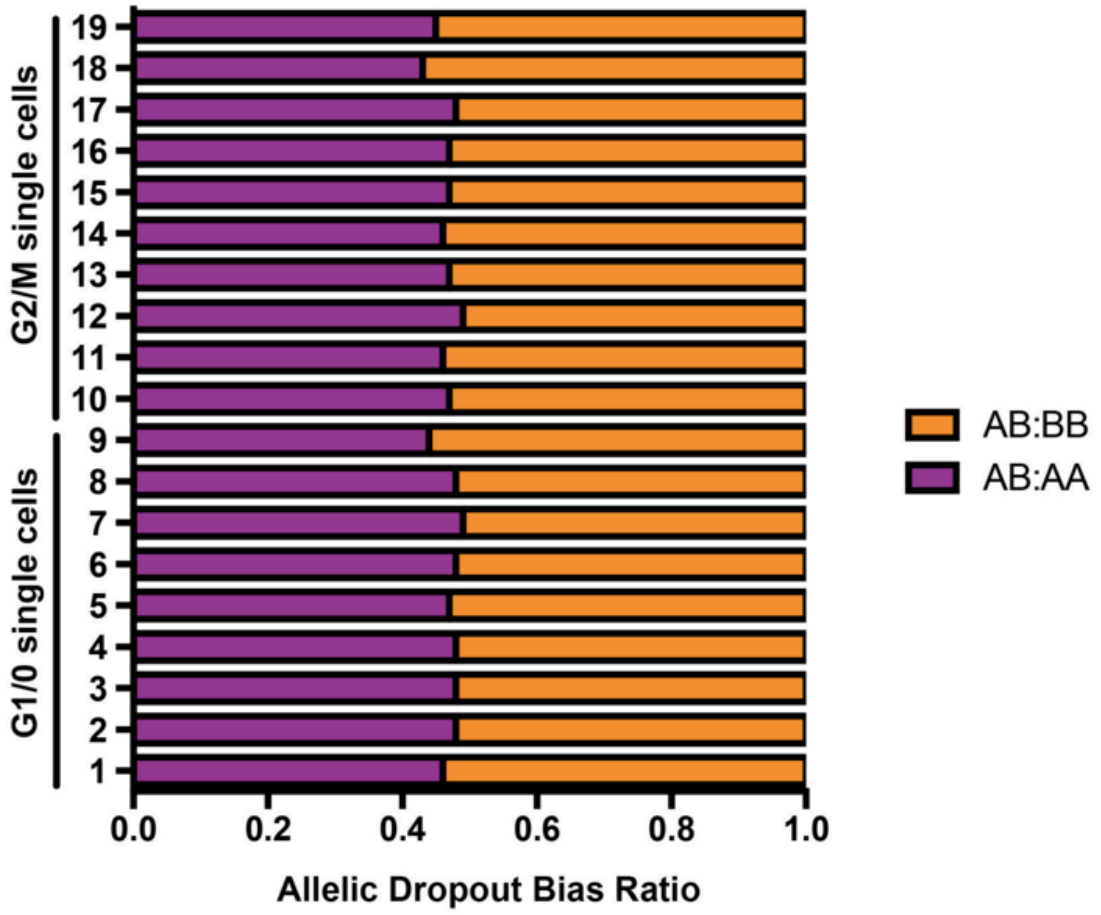

Figure 29 - Distribution of Allelic Dropout Bias

Cells are slightly biased toward AB -> BB dropout event, compared to AB -> AA events, for both G1/0 and G2/M cells.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

distribution of FP errors that occur at difference sites in single cells, allelic dropout errors sometimes occurred at recurrent position in multiple single cells. On average we observed that 2.55 cells out of 19 single cells shared a recurrent allelic dropout event at the same nucleotide position. These regions are important to note in single-cell studies and are showed be filtered, since they can misinterpreted as biological variation in SNVs prevalence, when in fact they are likely to be technical errors.

4.2.4 Measuring Detection Efficiency

We calculated the detection efficiencies, to measure the proportion of the SNvs and indels that were successfully detected in each single fibroblast cell exome. For SNVs, we detected 92.37% (±0.008, sem) of the variants in the single cells (mean = 32,369/34,982) in the G2/M cells, and 86.71% (±0.012, sem) in the G1/0 cells (mean = 25,753/29,549). (Figure 30) In comparison, previous studies using MALBAC reported detection efficiencies of only 76% for SNVs. An alternative approach is to calculate the SNV detection efficiency at all variant sites in the reference, regardless of the coverage depth in the single-cell and population sample. This calculation results in a detection efficiency for SNVs of 60.64% for G1/0 cells and 76.22% for G2/M cells. We also calculated the detection efficiency for indels, which is 85.60% (±0.007 SEM) for G2/M cells (mean = 2,448/2,856), and 82.11% (±0.009 SEM) for G1/0 cells (mean = 1,926/2,336) (Figure 31). To our knowledge, this is the first report showing that indels can accurately be detected in the genomes of single mammalian cells.

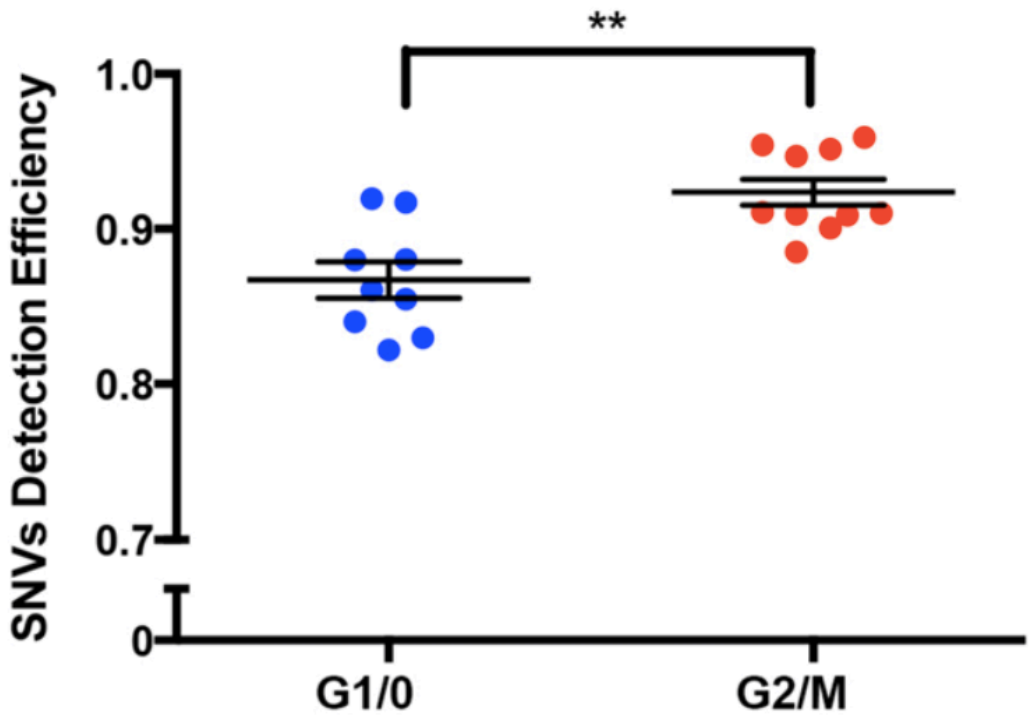

Figure 30 - Detection Efficiency for SNVs in Single Cells

G1/0 cells show 86.71% of detection efficiency in SNVs for G1/0 cells and 92.37% in G2/M cells.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

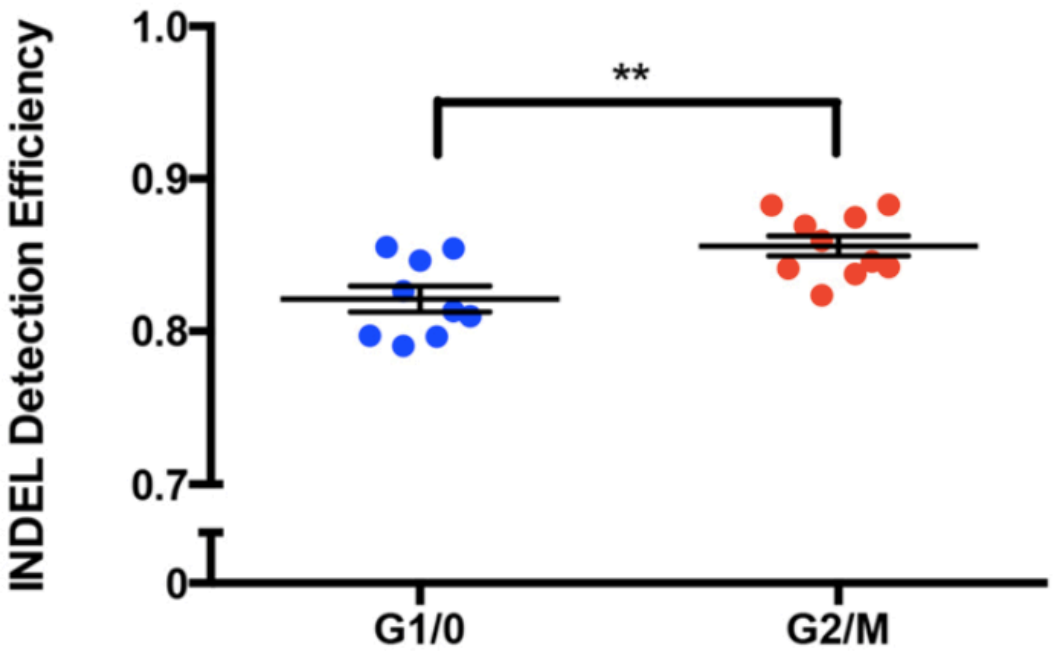

Figure 31 - Detection Efficiency for Indels in Single Cells

G1/0 cells show 82.11% of detection efficiency in indels for G1/0 cells and 85.60% in G2/M cells.

(Modified and reproduced from Marco L. Leung, Yong Wang, Jill Waters & Nicholas E Navin. SNES: Single Nucleus Exome Sequencing. *Genome Biology*. 16:55. 03/2015. PMID:25853327.

Copyright permission is not required since *Genome Biology* states, “the authors retain copyright of their article.”)

4.3 Discussion

This chapter has described the methodology and detailed technicality of SNES, a method that can achieve high coverage (96%) data from the exome of a single mammalian cell. From these data, we show that we can accurately detect SNVs and indels at base-pair resolution. The technical performance in coverage improvement is due to multiple factors, including an improved phi29 polymerase, time-limited isothermal amplification and the use of a 22-chromosome qPCR panel to eliminate cells with poor WGA performance prior to exome capture and sequencing.

In this chapter, we have improved the sequencing quality of the single cells and decreased the error rates. In the Xu *et al* study, they have reported very high ADR of 43.09% in single-cell exome sequencing data, whereas we have decreased ADR to 21.52%.[76] In the Hou *et al* study using MALBAC, they reported detection efficiencies of only 76% for SNVs, whereas we can achieve up to 92.37%. We have eliminated the NUC-Seq requirement for Tn5 transposase for library construction, which can introduce integration biases in the human genome and lead to coverage non-uniformity. Moreover, SNES protocol eliminates commercial kits for cell isolation, WGA and library construction, thereby reducing the cost of generating a single-cell library to approximately $30 per cell. In other words, SNES demonstrates superior single cell data quality, compared to other existing single cell genome sequencing methods.

CHAPTER FIVE – HIGHLY MULTIPLEXED TARGETED DNA SEQUENCING

Chapter 5 – Highly Multiplexed Targeted DNA Sequencing

Content of this chapter is based on: Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

5.1 Introductions and Rationale

To thoroughly understand intratumor heterogeneity, many single cells are required to study in order to accurately survey the diverse cell population in human cancers. Unlike single cell copy number profiling, whole genome or exome sequencing requires more sequencing data for analysis, thus it is difficult to multiplex many single cell libraries into one lane.

To address this issue, we have further refined SNES by using DNA barcoding to multiplex 48-96 single cells into single sequencing reactions to further increase throughput and to reduce costs of SCS. This is achieved by performing targeted capture on a panel of 201 cancer-associated genes, resulting in high coverage depth and reducing the cost of sequencing. In this chapter, we describe the changes of our method from **Chapter 4**, as well as the metrics generated by this updated method.

5.2 Results

5.2.1 Reduction of Captured Exome Region to T200 Panel

To increase the number of single cell samples in a flowcell lane, we compensate by decreasing the captured region from exome to 201 genes. We adopt the gene panel from the Chen *at al* study.[80,88] The genes are listed in Table 4. These genes are found mutated in 5% or more of the samples across all cancers, and 3% or more of the samples in 1 specific cancer types when at least 50 samples had been available. These genes can also be targeted by drugs that were commercially available, in clinical trails or under late-stage preclinical development. Several large genes previously shown mutated in cancer but with no direct clinical implications were not included, such as titin (*TTN*), Wolf-Hirschhorn syndrome (*WGHSC1*, also known as *NSD2*), and microtubule-actin crosslinking factor 1 (*MACF1*).[80] This panel covers 4,875 exons, spanning 938,607 bases.

5.2.2 Metric Performance

To establish technical error rates and metrics for this protocol, we applied our method to an isogenic breast cancer cell line (MDA-MB-231) to sequence 46 single cells and two matched bulk populations. We constructed 48 barcoded libraries and pooled together 46 single cells libraries and two matched population samples into a single reacton for targeted capture using the T200 panel of 201 cancer-associated genes. (Table 4) The pooled libraries were sequenced on a single lane on a HiSeq 2000 system (Illumina) at 100-bp paired-end cycles. Our samples showed an average coverage depth of 255x (SEM = 23.54) and

ABL1	BRCA2	CSMD2	EZH2	GNAS	KDM6A	MLL3	PAX5	PTPN11	STK11
ACVR1B	CARD11	CSMD3	FAM123B	HDAC9	KDR	MPL	PBRM1	PAD51	SYK
ADAMTS12	CASP8	CTNNB1	FAM135B	HEATR7B2	KIT	MSH2	PCDH15	RAF1	SYNE1
AKAP3	CBL	CYLD	FAT3	HGF	KRAS	MSH6	PCLO	RB1	SYNE2
AKT1	CD19	CYP2C19	FBXW7	HMCN1	LAMA1	MTOR	PDGFRA	RELN	TBC1D4
ALK	CDH1	DAXX1	FGFR1	HNG1A	LPHN3	MYD88	PDGFRB	RET	TET2
APC	CDH10	DDR1	FGFR2	HNG1B	LRP1	NAV3	PIK3CA	RIMS2	TGFb1
AR	CDH11	DDR2	FGFR3	HRAS	LRP1B	NCOR1	PIK3CG	RNF213	TGFBR2
ARAF	CDK4	DNMT3A	FGFR4	HYDIN	LRP2	NF1	PIK3R1	RUNX1	TNFAIP3
ARID1A	CDK8	EGFR	FLG	IDH1	MAP2K1	NF2	PIKFYVE	RUNX1T1	TOP1
ASXL1	CDKN2A	ELN	FLT1	IDH2	MAP2K4	NFKB2	PKHD1	RYR2	TOP2A
ATM	CEBPA	EML4	FLT3	IGF1R	MAP3K1	NOTCH1	PKHD1L1	SETD2	TP3
ATR	CHEK1	EP300	FL4	IKZF1	MAP3K4	NOTCH2	PPP1R3A	SMAD4	TSC1
ATRX	CHEK2	EPHA3	FOXL2	IL6R	MDN1	NOTCH3	PPP2R1A	SMARCA4	TSC2
AURKA	COL14A1	ERBB2	GABRA6	IRAS1	MECOM	NOTCH4	PPP2R4	SMARCB1	TSHR
AURKB	CPAMD8	ERBB3	GABRB3	ITGA4	MEN1	NPM	PRDM1	SMO	USH2A
BAI3	CREBBP	ERCC3	GATA1	JAK1	MET	NRAS	PRSS1	SOS1	VHL
BAP1	CRIPAK	ERCC4	GATA3	JAK2	MITF	NSD1	PTCH1	SPEN	WHSC1
BRAF	CSF1R	ERCC5	GNA11	JAK3	MLH1	PALB2	PTEN	SPOP	WT1
BRCA1	CSMD1	ETV5	GNAQ	KCNB2	MLH2	PAPPA2	PTK2	SPTA1	ZNF238
									ZNG536

Table 4 - Genes Targeted by the T200 Panel

These genes are selected in the Chen *at al* study and found mutated in 5% or more of the samples across all cancers, and 3% or more of the samples in 1 specific cancer types when at least 50 samples had been available. These genes can also be targeted by drugs that were commercially available, in clinical trails or under late-stage preclinical development.

coverage breadth of 85% (SEM = 0.02%). (Figure 32) Uneven pooling can lead to occasional samples with low coverage, which should be removed from downstream analysis (for example, cell number 46 in Figure 32). The average on-targeted performance for this SCS data set in the capture regions was determined to be 65.03%.

Next, using the variants detected in the isogenic population samples. We calculated the technical error rates for each single cell at sites at which both samples ad sufficient coverage depth. The mean ADR for the single-cell data was 13.68% (SEM = 1.9%; Figure 33). We also calculated the detection efficiency of SNVs in regions in which sufficient coverage (≥10x) was found in both the population and single cell data sets. Our analysis identified an SNV detection efficiency of 82.80% (SEM = 1.9; Figure 34) We calculated the mean false positive error rate to be $4.98e^{-5}$ (SEM = $4.175e^{-6}$; Figure 35). This error rate is drastically reduced (squared) by calling mutations concurrently in two or more single cells.

5.3 Discussion

In this chapter, to increase the number of multiplexed samples, we decrease the region covered by the target capture from exome to 200 genes. These 201 genes are chosen based on their clinical relevance and frequent occurrence in all cancer. By sequencing only 201 genes instead of exome, the number of variants detected will decrease. There may be mutations with significant consequences that are not included in this platform.

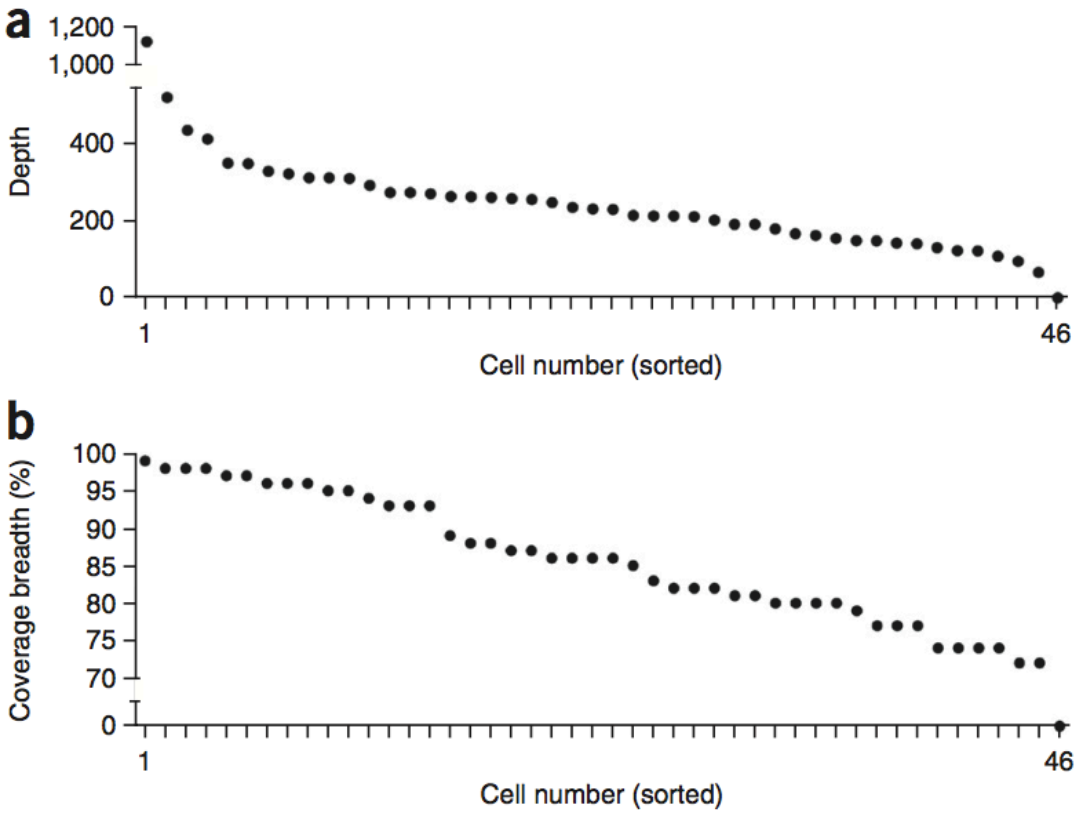

Figure 32 - Coverage Metrics for Single Cells

46 single cells were multiplexed and sequenced at average coverage depth of 255x (a) and coverage breadth of 85% (b).

(Modified and reproduced from Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

Copyright permission is not required since Nature journal policy states “author retains the copyright to the published materials.”)

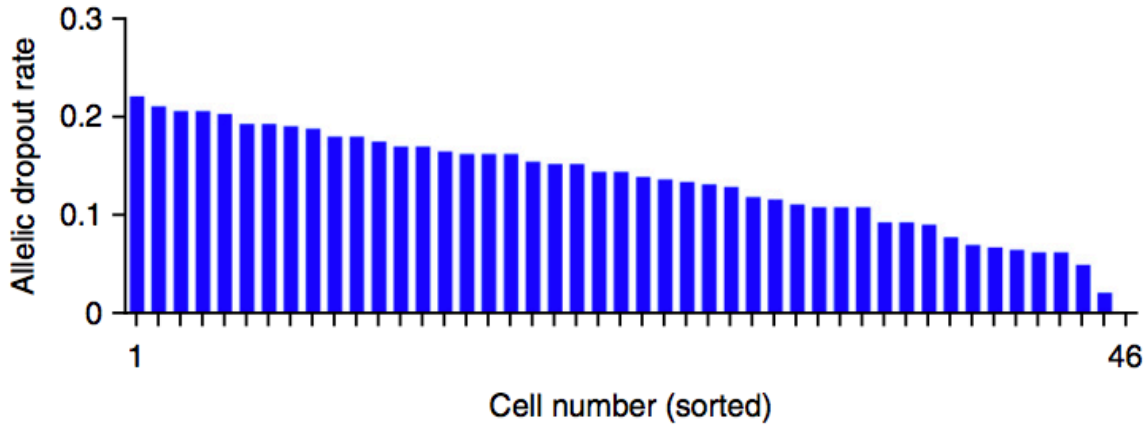

Figure 33 - Single Cell Allelic Dropout Rate

The average allelic dropout rate for 46 single cells was 13.68%.

(Modified and reproduced from Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

Copyright permission is not required since Nature journal policy states “author retains the copyright to the published materials.”)

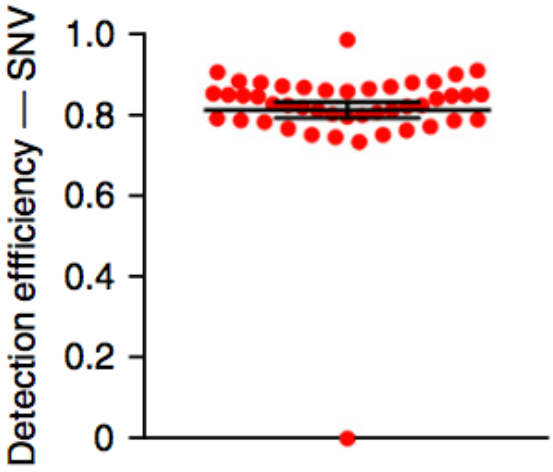

Figure 34 - Detection Efficiency for Single Nucleotide Variants

The average detection efficiency for SNVs is 82.80%.

(Modified and reproduced from Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

Copyright permission is not required since Nature journal policy states “author retains the copyright to the published materials.”)

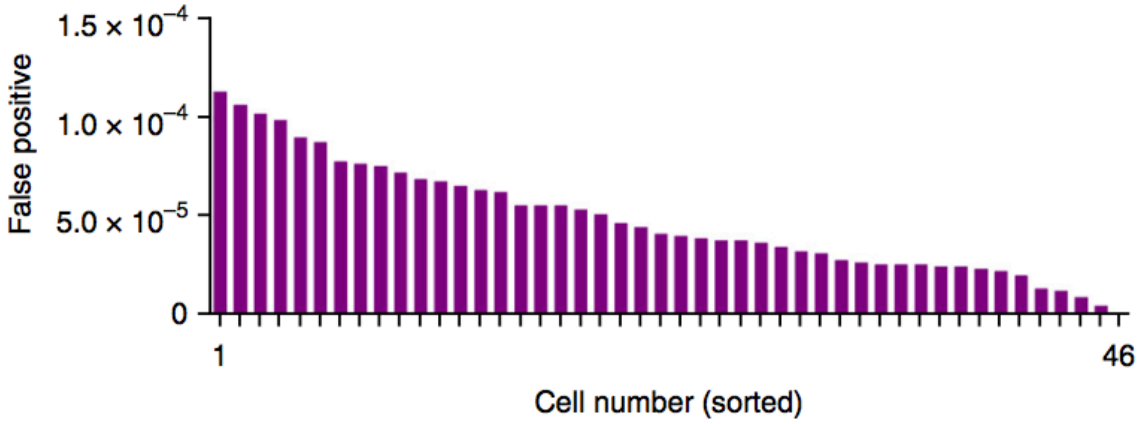

Figure 35 - False Positive Rate

The average false positive error rate for single cells is $4.87e^{-5}$.

(Modified and reproduced from Marco L. Leung, Yong Wang, Charissa Kim, Ruli Gao, Emi Sei & Nicholas E Navin. Highly-Multiplexed Targeted DNA Sequencing of Single Nuclei. *Nature Protocols*. 2016 Feb;11(2):214-35. doi: 10.1038/nprot.2016.005. Epub 2016 Jan 7.

Copyright permission is not required since Nature journal policy states “author retains the copyright to the published materials.”)

Collectively, these data show that high-coverage breadth data and high detection efficiencies can be obtained using the protocol to perform highly multiplexed single-cell targeted DNA sequencing. By multiplexing up to 96 samples in one sequencing lanes, more single cells can be observed and high level of clonality within a tumor can be detected. We will further discuss how many single cells are needed to detect intratumor heterogeneity in a later section (**7.1.3 Determining the number of single cells required for sampling**).

CHAPTER SIX - TRACING METASTATIC LINEAGE IN COLORECTAL CANCER USING SINGLE CELL SEQUENCING

Chapter 6 – Tracing Metastatic Lineage in Colorectal Cancer Using Single Cell Sequencing

6.1 Introductions and Rationale

As stated in **Chapter 1**, patients with CRC metastasis have worse prognosis than those without metastasis. Understanding how cancer genomes evolve during CRC metastasis may provide valuable insights on treating CRC metastasis. Recent studies have attempted to investigate the genomic diversity and metastasis among hundreds of patients using next generation sequencing (NGS). The TCGA study had found genes (*APC*, *KRAS*, *NRAS*, *TP53)* that are frequently mutated in CRC, as well as genes that are mutated at low frequency, such as *ARID1A*, *SOX9* and *FAM123B*.[11] Moreover, it was found that, in non-MSI CRC, primary and metastatic tumors had high concordance of mutational profiles in most patients.[89,90] This suggests that CRC follows the late-dissemination model, in which the primary tumor has progressed for a long period of time before tumor cells metastasize to remote organs. However, it is difficult to detect rare clones in heterogeneous tumors using conventional NGS methods. There might be metastatic subpopulations that exist at low clonal frequency and cannot be detected by conventional NGS method.

Here, we used SCS methods (as described in **Chapter 4** and **5**) to sequence single cells from primary tumors and metastases from two CRC patients. Our data identified a large number of nonsynonymous mutations that evolved in the root nodes during the earliest stages of primary tumor evolution and were maintained in all single cells during the clonal expansion of the tumor

mass. We explored the clonality of these tumors from bulk and single cell sequencing. Using the single cell data, we construct phylogenetic trees, which reveals branched evolution at both organ sites.

6.2 Results

6.2.1 Bulk Exome Sequencing Analysis of Two Colorectal Cancer Patients

For our study, we selected two colorectal cancer male patients (CO5 and CO8). Both patients have invasive differentiated adenocarcinoma of the colon without microsatellite instability. Both patients are presented with liver metastases. First, we isolated the nuclei from these tumors and stained with DAPI before we flow-sorted them based on ploidy distribution. (Figure 36) We gated and flow-sorted the bulk and single cells of each diploid and aneuploid population from each tumor. We performed exome sequencing on these four populations, as well as the matched normal tissues to filter germline variants, with high coverage performance (average coverage depth = 75.5x, average coverage breadth = 0.9733). (Table 5) We detected 90 nonsynonymous mutations for CO5 and 107 for CO8. These mutations include *KRAS*, *NRAS*, *APC* and *TCF7L2*, and they have been previously reported in the TCGA colorectal cancer study.[11] (Table 6,7)

Consistent with previous studies, we found that the majority of mutations were common in the primary and metastatic tumors.[89,90] There are 80 and 68 common mutations in CO5 and CO8, respectively, with minor number of site-specific mutations. (Figure 37) We plotted the primary mutation frequency against the metastasis mutation frequency, and found that *APC*, *KRAS*, and

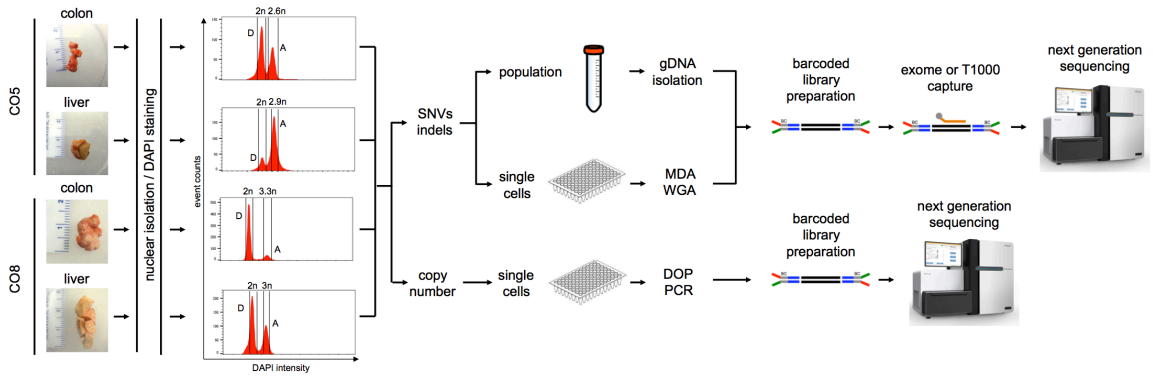

Figure 36 - Experimental Workflow

Each tumor is gated and flow-sorted by ploidy distribution. For SNVs and indels detections, single cells are amplified using multiple-displacement-amplification with Φ29 polymerase, while degenerate-oligo-primed PCR is used for single cell copy number detection. For SNVs and indels, libraries are captured using either exome or T1000 hybridization. For copy number profiling, single cells are amplified using DOP-PCR and multiplexed for Illumina sequencing.

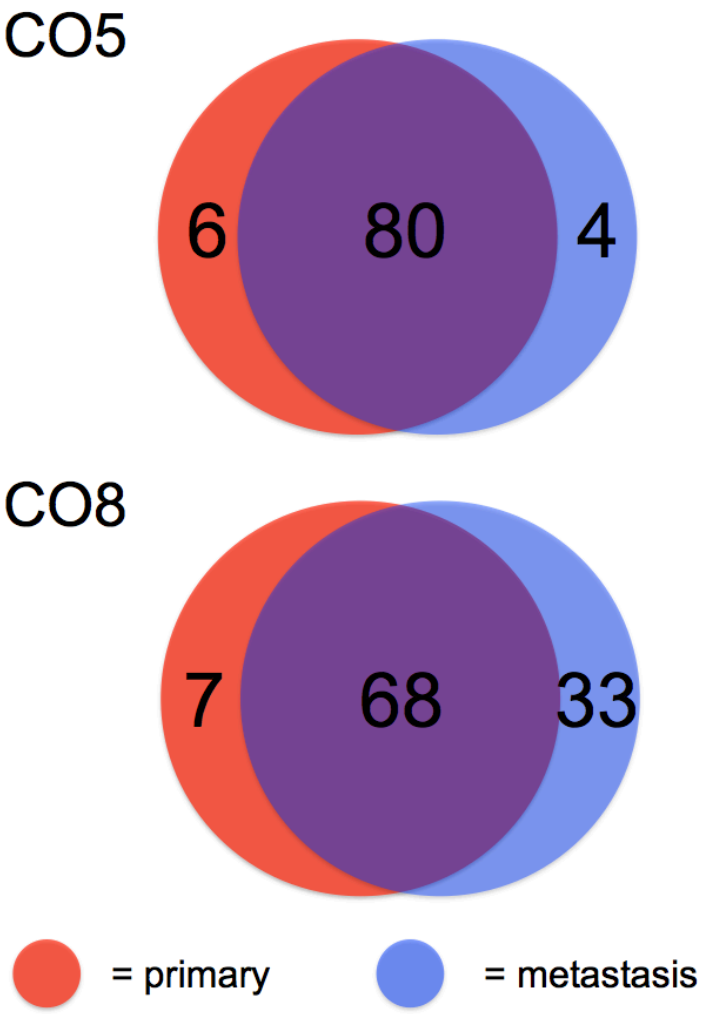

Figure 37 - Nonsynonymous Mutations Detected in CRC Tumors

Venn diagrams demonstrate the number of mutations that are common and specific to each tumor.

CO5	Depth (x)	Breadth
Normal colon	59	0.9596
Normal liver	63	0.9417
Primary Diploid	63	0.9684
Primary Aneuploid	80	0.9819
Metastatic Diploid	49	0.9744
Metastatic Aneuploid	70	0.9786

CO8	Depth (x)	Breadth
Normal colon	111	0.9812
Normal liver	83	0.9776
Primary Diploid	93	0.9805
Primary Aneuploid	93	0.98
Metastatic Diploid	46	0.9756
Metastatic Aneuploid	96	0.9798

Table 5 - Coverage Metrics for Bulk Exome Sequencing

The average depth and breadth for CO5 is 64x and 0.9674, respectively. The average depth and breadth for CO8 is 87x and 0.9791, respectively.

Chromosome	Position	Gene Name	Reference	Variant
chr9	95784648	FGD3	G	A
chr10	131641447	EBF3	G	A
chr19	50214113	CPT1C	C	T
chr19	5244392	PTPRS	C	T
chr15	42438030	PLA2G4F	C	T
chr8	27779273	SCARA5	C	T
chr16	511409	RAB11FIP3	G	T
chr2	27552346	GTF3C2	G	A
chr17	3772840	CAMKK1	C	T
chr7	75050957	POM121C	C	T
chrX	48650491	GATA1	C	A
chr9	116931070	COL27A1	G	A
chrX	7811288	VCX	C	G
chr10	114911615	TCF7L2	C	A
chr22	39629508	PDGFB	T	C
chr19	2216629	DOT1L	C	T
chr7	36492153	ANLN	G	T
chr11	111249887	POU2AF1	G	A
chr12	42512910	GXYLT1	T	A
chr15	41056381	GCHFR	G	A
chr1	181686322	CACNA1E	G	A
chr7	75050891	POM121C	T	C
chr16	16103671	ABCC1	C	A
chr17	27383288	PIPOX	A	C
chr17	39978517	FKBP10	C	T
chr5	112175303	APC	C	T
chr17	72739280	RAB37	G	A
chr11	6470323	TRIM3	C	A
chr1	197111575	ASPM	G	T
chr12	25398285	KRAS	C	A
chr11	67012738	KDM2A	C	T
chr4	109672135	ETNPPL	C	T
chr19	58320385	ZNF552	C	T
chr1	111957412	OVGP1	G	A
chr20	43851625	SEMG2	C	G
chr9	43625849	SPATA31A6	C	G
chr4	187455223	MTNR1A	G	A
chr4	114278764	ANK2	C	T
chr2	114512750	SLC35F5	G	A
chr19	9048227	MUC16	A	G
chr4	96761627	PDHA2	C	T
chr2	152515652	NEB	T	G
chr18	61570307	SERPINB2	A	G
chr11	111177170	COLCA2	G	A
chr19	30313229	CCNE1	C	T
chr5	169535601	FOXI1	G	T
chr3	38739105	SCN10A	C	T
chr17	56270739	EPX	C	A
chr17	72832510	TMEM104	C	T
chr11	72945731	P2RY2	C	T
chr17	12905805	ELAC2	G	A
chr4	2691303	FAM193A	A	G
chr1	27332541	FAM46B	C	T
chr9	137620520	COL5A1	C	T
chr10	81697853	SFTPD	C	T
chr16	80718449	CDYL2	G	A
chr10	82348482	SH2D4B	C	A
chr15	86838537	AGBL1	C	T
chr3	187387985	SST	C	T
chr5	149677523	ARSI	C	T
chr5	141237011	PCDH1	G	A
chrX	48207025	SSX3	C	T
chr17	42854546	ADAM11	G	A

chr5	137766019	KDM3B	C	T
chr18	22807094	ZNF521	T	C
chr1	6681651	PHF13	G	A
chrX	38135983	RPGR	C	T
chr5	54423155	CDC20B	G	A
chrX	135958730	RBMX	C	A
chr7	5415673	TNRC18	G	A
chr7	142458526	PRSS1	A	G
chr17	39394674	KRTAP9-8	A	G
chr2	131797751	ARHGEF4	C	T
chr1	36564618	COL8A2	C	T
chr13	109792732	MYO16	C	A
chr5	156279	PLEKHG4B	G	A
chr2	166929996	SCN1A	C	T
chr22	24581995	SUSD2	G	A
chr5	140594292	PCDHB13	A	C
chr22	20460526	RIMBP3	C	T
chrX	101912757	GPRASP1	A	T
chr19	15292500	NOTCH3	G	T
chr1	152279527	FLG	T	C
chr15	44038842	PDIA3	C	G
chr4	170042033	SH3RF1	C	G
chr11	62294315	AHNAK	G	C
chr5	173035291	BOD1	G	A
chr7	98554034	TRRAP	A	G
chr4	106755675	GSTCD	A	G
chr11	130785060	SNX19	A	T

Table 6 – CO5 Exome Bulk Sequencing Mutations

The table lists the mutations detected in the CO5 tumors using exome sequencing. Black represents mutations detected in both primary and metastatic tumors. Blue represents primary-specific mutations and red represents metastatic mutations.

Chromosome	Position	Gene name	Reference	Variant
chr16	10525156	ATF7IP2	G	T
chr8	51465694	SNTG1	G	T
chr3	38648271	SCN5A	C	A
chr5	666171	TPPP	G	A
chr6	56883251	BEND6	G	T
chr1	46290133	MAST2	T	G
chr4	114279178	ANK2	G	T
chr20	60585112	TAF4	G	A
chr10	84718709	NRG3	C	T
chr8	59851979	TOX	T	C
chr3	136076689	STAG1	C	T
chr7	32909384	KBTBD2	T	C
chr1	228504485	OBSCN	C	T
chr2	215880335	ABCA12	G	A
chr11	117299235	DSCAML1	T	G
chr12	58144548	CDK4	C	A
chr4	186545169	SORBS2	C	T
chr16	15917267	MYH11	G	A
chr8	77767363	ZFHX4	C	T
chr12	9254240	A2M	C	T
chr1	103471858	COL11A1	A	G
chr8	106814316	ZFPM2	A	G
chr1	3645901	TP73	G	A
chr19	8613192	MYO1F	G	T
chr1	217975125	SPATA17	A	C
chr5	16694700	MYO10	G	A
chr7	48563978	ABCA13	G	A
chr11	117279728	CEP164	G	T
chr11	44069747	ACCSL	C	T
chr12	112701998	HECTD4	G	A
chr11	66392695	RBM14	G	A
chr6	50810945	TFAP2B	C	T
chr15	28520057	HERC2	G	A
chr20	5283324	PROKR2	G	T
chr5	112164646	APC	G	T
chr10	27687804	PTCHD3	C	T
chr11	119053871	NLRX1	G	A
chr11	48373961	UNKNOWN	A	T
chr10	52603881	A1CF	T	G
chr8	59059734	FAM110B	C	A
chr1	115258747	NRAS	C	A
chr19	52130463	SIGLEC5	G	C
chr15	41810233	RPAP1	A	T
chr22	28193785	MN1	G	T
chr2	179426759	TTN	C	A
chr11	32635763	CCDC73	G	A
chr9	131483555	ZDHHC12	G	A
chr17	17129519	FLCN	G	A
chr11	65978634	PACS1	C	A
chr21	30934019	GRIK1	G	A
chr6	163956109	QKI	C	G
chr2	37105068	STRN	G	A
chr20	35444571	SOGA1	T	C
chrX	133700173	PLAC1	T	G
chr2	226447238	NYAP2	G	A
chr7	48391820	ABCA13	C	T
chr12	75601447	KCNC2	C	T
chr11	120811150	GRIK4	T	A
chr2	167145040	SCN9A	T	G
chr1	205631135	SLC45A3	C	T
chr1	152275373	FLG	C	T
chrX	17819893	RAI2	G	T
chr7	94293611	PEG10	G	A

chr6	7374272	CAGE1	C	A
chr5	112175328	APC	C	A
chr10	49929315	WDFY4	G	A
chr15	43574260	TGM7	G	A
chr17	7577548	TP53	C	T
chr6	90422940	MDN1	C	T
chrX	32536138	DMD	T	G
chr15	56122103	NEDD4	C	G
chr22	40417962	FAM83F	T	A
chr21	43691270	ABCG1	C	G
chr11	101771267	ANGPTL5	C	A
chr10	15821128	FAM188A	C	G
chr19	9071750	MUC16	T	G
chr1	225156539	DNAH14	T	C
chr6	153345483	RGS17	A	C
chr5	178140358	ZNF354A	A	T
chr1	211192300	KCNH1	A	C
chr5	106762962	EFNA5	G	T
chr1	16258997	SPEN	G	C
chr21	19638334	CHODL	A	C
chr2	21228452	APOB	A	C
chr22	26898017	TFIP11	T	A
chr2	11716651	GREB1	G	C
chrX	154157378	F8	C	G
chr22	46930524	CELSR1	C	T
chr5	127866348	FBN2	G	T
chr12	25260947	LRMP	T	A
chr14	20711786	OR11H4	T	C
chr12	102811648	IGF1	C	A
chr2	128624549	AMMECR1L	A	G
chr5	140573225	PCDHB10	T	A
chr6	18197451	KDM1B	A	C
chr16	31193877	FUS	C	T
chr14	45639916	FANCM	G	C
chr7	92120681	PEX1	C	T
chr5	38943059	RICTOR	T	G
chr4	119035963	NDST3	A	C
chr1	186097285	HMCN1	G	A
chr8	55539448	RP1	A	C
chrX	36329023	CXorf30	C	T
chrX	114398248	LRCH2	T	G
chr11	45671752	CHST1	G	T
chr15	22958342	CYFIP1	A	C
chr4	158224914	GRIA2	A	G
chr9	90321594	DAPK1	C	T

Table 7 – CO8 Exome Bulk Sequencing Mutations

The table lists the mutations detected in the CO8 tumors using exome sequencing. Black represents mutations detected in both primary and metastatic tumors. Blue represents primary-specific mutations and red represents metastatic mutations.

NRAS were present at high frequency in both primary and metastatic tumors. (Figure 38) This suggests that these mutations occurred early during the tumorigenesis, thus most tumor cells had acquired these mutations.

Next, we performed amplicon-sequencing on metastasis-specific mutations of CO5 and CO8 in the primary tumors. This is to confirm whether these mutations exist at low frequency in the primary tumor. We sequence 3 and 18 metastasis-specific mutations in CO5 and CO8, respectively. We did not find any mutations that have a significantly higher read count than the normal tissue; hence the metastasis-specific mutations are not present in the primary tumor at low frequency. (Figure 39) (Table 8)

6.2.2 Single Cell SNVs Analysis

To resolve the intratumor heterogeneity and trace the metastatic lineage, we sequenced single tumor cells from both patients. Using our previous published method, we amplified the single cell genome using multiple displacement amplification and performed barcoded library construction. (Figure 36) We performed targeted sequencing on diploid and aneuploid cells by capturing a 4 Mb region, spanning 1000 genes.

We sequence 186 single cells from the primary and metastatic tumors for each patient. The average coverage depth is 137x and average coverage breadth is 0.92. (Figure 40 and 41) Single cell data are processed using our pipeline described in Chapter 2. (Figure 3) Furthermore, we applied additional filtering steps to select for variants. (Figure 42 and 43)

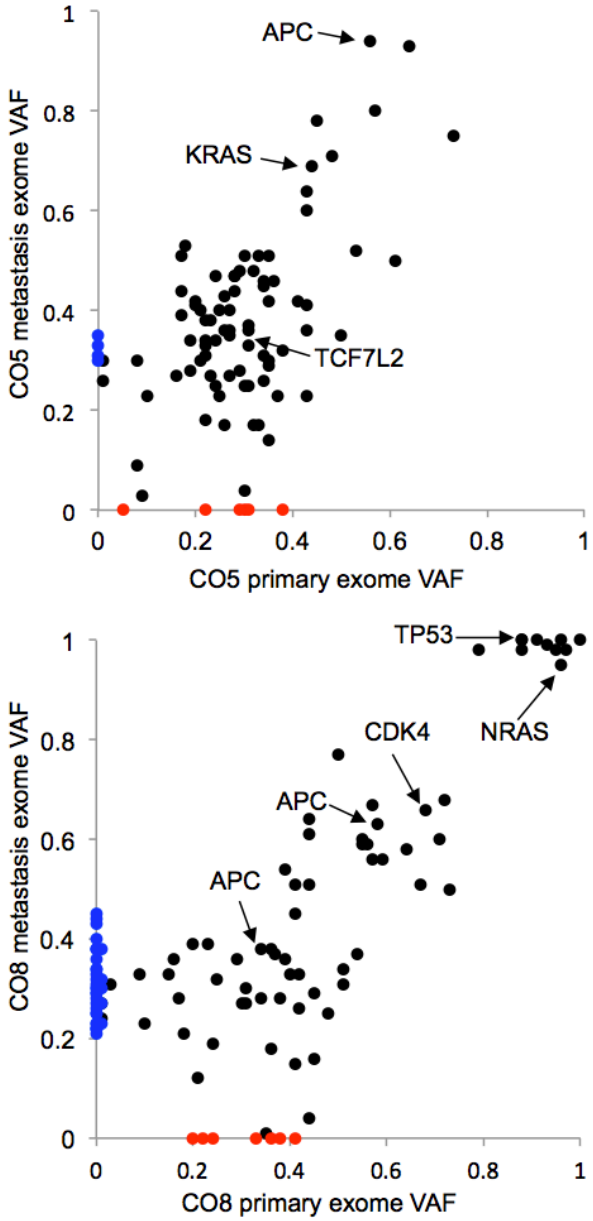

Figure 38 - Variant Allele Frequency of Primary and Metastatic Tumors

The primary bulk exome variant allele frequency is plotted against those of metastasis. Red dots represent mutations that are specific to primary. Blue dots represent mutations that are specific to metastasis.

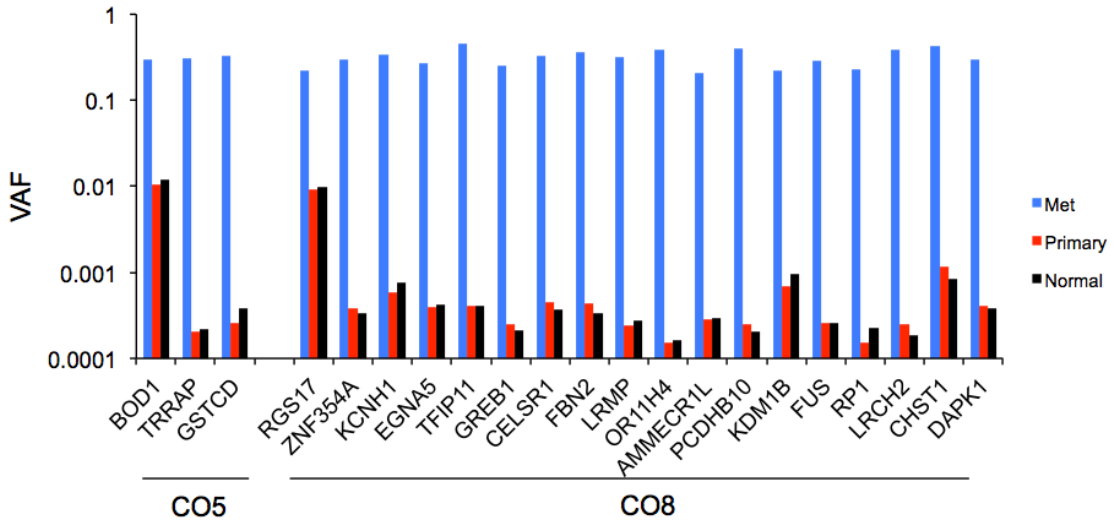

Figure 39 - Amplicon Sequencing of Metastasis-specific Mutations

150bp regions of metastasis-specific mutations were flanked by PCR primers and amplified in the genomic DNA. Sequencing read counts of the reference and variant allele were calculated and compared. Blue represents the metastasis. Red represents the primary tumor and black represents the normal tissue.

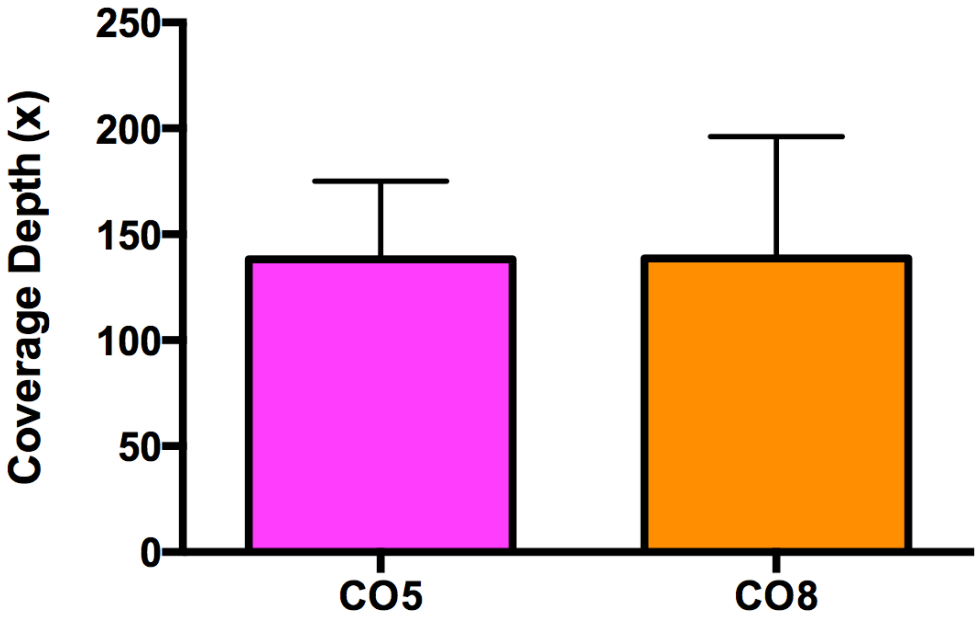

Figure 40 - Coverage Depth for Single Cells Data

The average coverage depth is 137.9x (SEM = 2.716) for CO5 and 138.6x (SEM = 4.215) for CO8.

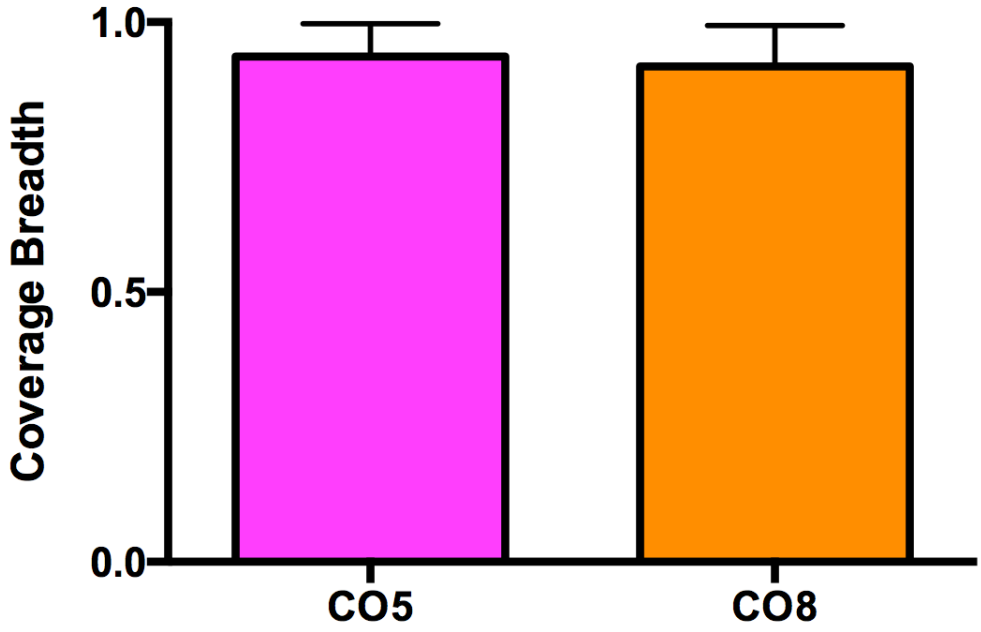

Figure 41 - Coverage Breadth for Single Cell Data

The average coverage breadth is 0.9363 (SEM = 0.0045) for CO5 and 0.9183 (SEM = 0.0055) for CO8.

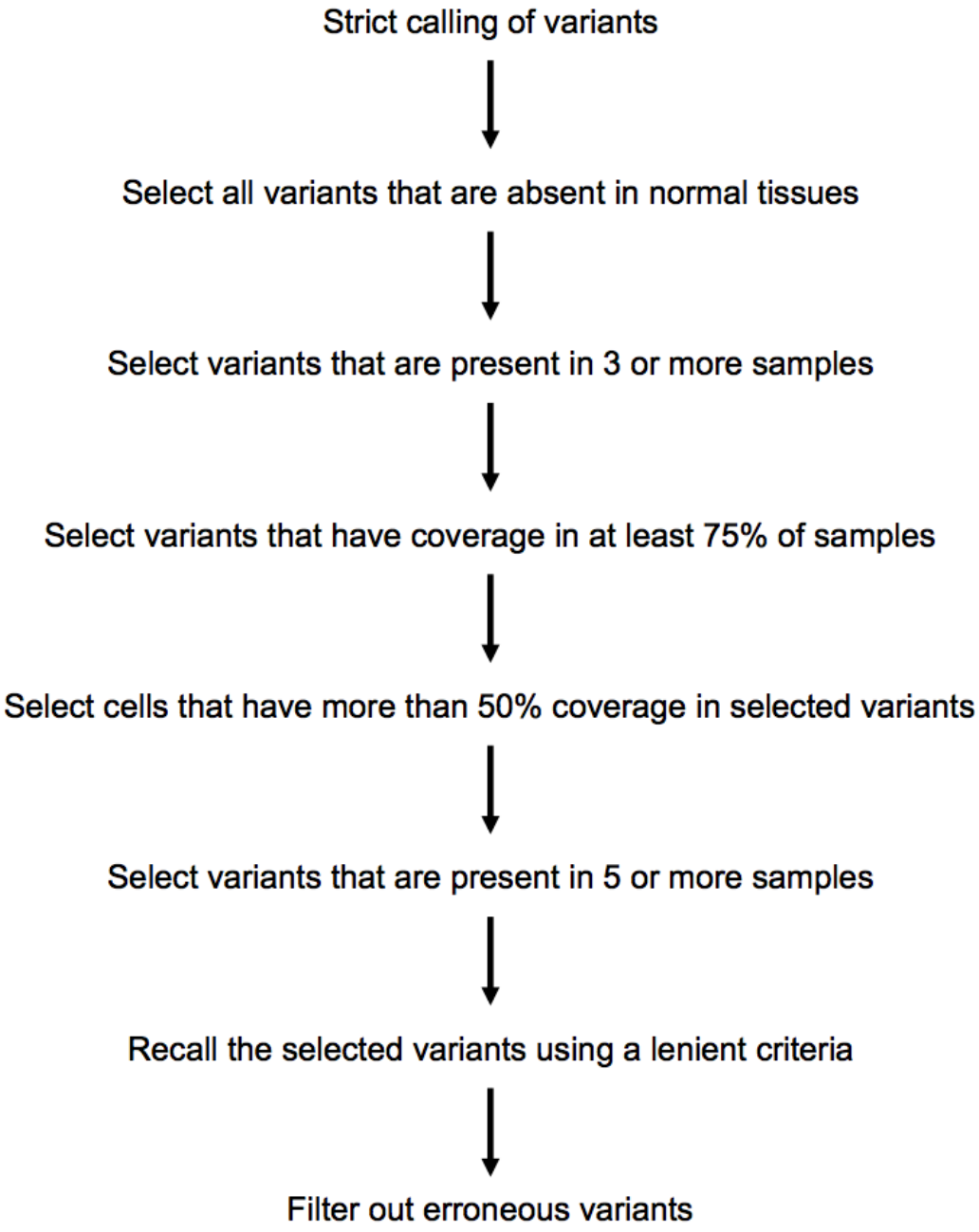

Figure 42 - Filtering Pipeline for CO5

Additional filtering steps were used to identify mutations in single cell in CO5.

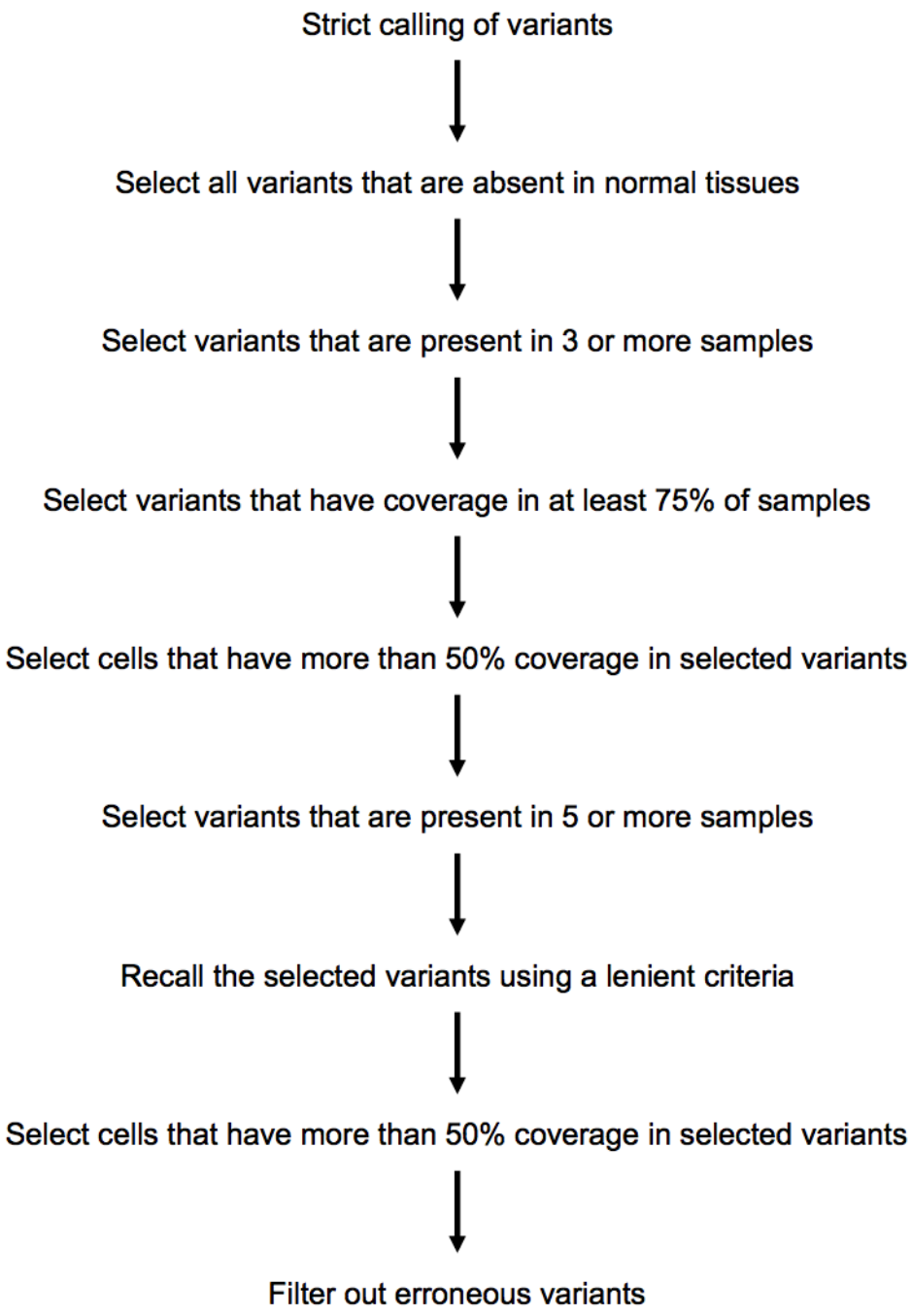

Figure 43 - Filtering Pipeline for CO8

Additional filtering steps were used to identify mutations in single cell in CO8.

						Met	normal					primary						
	chr	pos	gene	ref	var	vaf	A	C	G	T	total	vaf	A	C	G	T	total	vaf
	5	173035291	BOD1	G	A	0.3	64	39	5106	146	5355	0.011951447	37	29	3322	102	3490	0.010601719
CO5	7	98554034	TRRAP	A	G	0.31	1445168	1008	318	302	1446796	0.000219796	1466151	496	303	348	1467298	0.000206502
	4	106755675	GSTCD	A	G	0.33	1500362	671	574	295	1501902	0.000382182	1632261	868	422	352	1633903	0.000258277
	6	153345483	RGS17	A	C	0.22	3613	36	11	37	3697	0.009737625	5247	49	32	52	5380	0.009107807
	5	178140358	ZNF354A	A	T	0.3	821913	246	283	278	822720	0.000337904	785945	259	260	298	786762	0.000378768
	1	211192300	KCNH1	A	C	0.34	1921380	1481	731	416	1924008	0.000769747	2023907	1193	631	325	2026056	0.000588829
	5	106762962	EGNA5	G	T	0.27	334	66	1290095	548	1291043	0.000424463	519	128	1838594	713	1839954	0.00038751
	22	26898017	TFIP11	T	A	0.45	592	929	1192	1448993	1451706	0.000407796	508	884	765	1232711	1234868	0.00041138
	2	11716651	GREB1	G	C	0.25	699	343	1636714	1299	1639055	0.000209267	560	303	1218420	965	1220248	0.00024831
	22	46930524	CELSR1	C	T	0.33	428	2007773	458	739	2009398	0.000367772	885	1412564	455	627	1414531	0.000443256
	5	127866348	FBN2	G	T	0.36	219	88	880028	295	880630	0.000334987	148	60	495614	215	496037	0.000433435
CO8	12	25260947	LRMP	T	A	0.32	399	703	1706	1445118	1447926	0.000275567	464	899	1619	1940018	1943000	0.000238806
	14	20711786	OR11H4	T	C	0.38	141	203	232	1246055	1246631	0.000162839	126	171	137	1125174	1125608	0.000151918
	2	128624549	AMMECR1L	A	G	0.21	1767704	838	518	749	1769809	0.000292687	2565317	1216	723	1247	2568503	0.000281487
	5	140573225	PCDHB10	T	A	0.4	308	662	1010	1521724	1523704	0.000202139	319	413	696	1299987	1301415	0.000245118
	6	18197451	KDM1B	A	C	0.22	2844026	2683	730	529	2847968	0.000942075	2771445	1922	653	453	2774473	0.000692744
	16	31193877	FUS	C	T	0.29	1868	2704367	610	692	2707537	0.000255583	2042	2658991	633	688	2662354	0.000258418
	8	55539448	RP1	A	C	0.23	1662490	377	402	241	1663510	0.000226629	530243	80	126	75	530524	0.000150794
	X	114398248	LRCH2	T	G	0.38	43	71	68	371793	371975	0.000182808	24	30	46	185557	185657	0.000247769
	11	45671752	CHST1	G	T	0.43	858	1482	1437757	1213	1441310	0.000841595	358	618	589642	678	591296	0.001146634
	9	90321594	DAPK1	C	T	0.3	2035	3215837	1613	1235	3220720	0.000383455	1670	2292808	1079	947	2296504	0.000412366

Table 8 - Amplicon Deep-Sequencing of Metastatic-Specific Mutations

150bp region around the metastatic-specific mutations were flanked, amplified and sequenced. Each base was counted and variant allele frequencies were calculated.

As shown in the CO5 mutation heatmap, cells are separated into three major clusters, the diploid cells, the primary aneuploid cells and the metastatic aneuploid cells. (Figure 44) For the aneuploid cells, the primary and metastatic cells share 10 mutations. There are five metastatic-specific mutations and 1 primary-specific mutation. For the diploid populations, cells are mostly wild-typed in sites where mutations are detected in aneuploid cells. As shown in the heatmap, there are occasional false positive variants detected in the diploid cells. (Figure 44) We have checked each individual variant detected in the diploid cells and can verify that they are false positive (in dark grey), which may be caused by poor base quality or poor mapping quality. However, we found three cells (PDD93, PD16, and PD41) that only have the *APC* mutation. We check the other variants sites for these three cells and they are all wild-typed.

For CO8, there are 14 shared mutations, with two primary-specific mutations and 14 metastatic-specific mutations. There are also three major clusters for the diploid cell, primary aneuploid cells and metastatic aneuploid cells. (Figure 45) Unlike CO5, there are three subpopulations (m1, m2 and m3) for the metastatic aneuploid cells. (Figure 45) For the diploid cells, we found a sub-cluster of cells (diploid cluster) that have five mutations (*NR3C2*, *ATR*, *ALK*, *EPHB6* and *SPEN*), which do not shared with the aneuploid cells. (Figure 45)

Next, using the SNVs data from CO5 and CO8, we built multi-dimensional scaling (MDS) plots. (Figure 46 and 47) We found that, for both patients, there

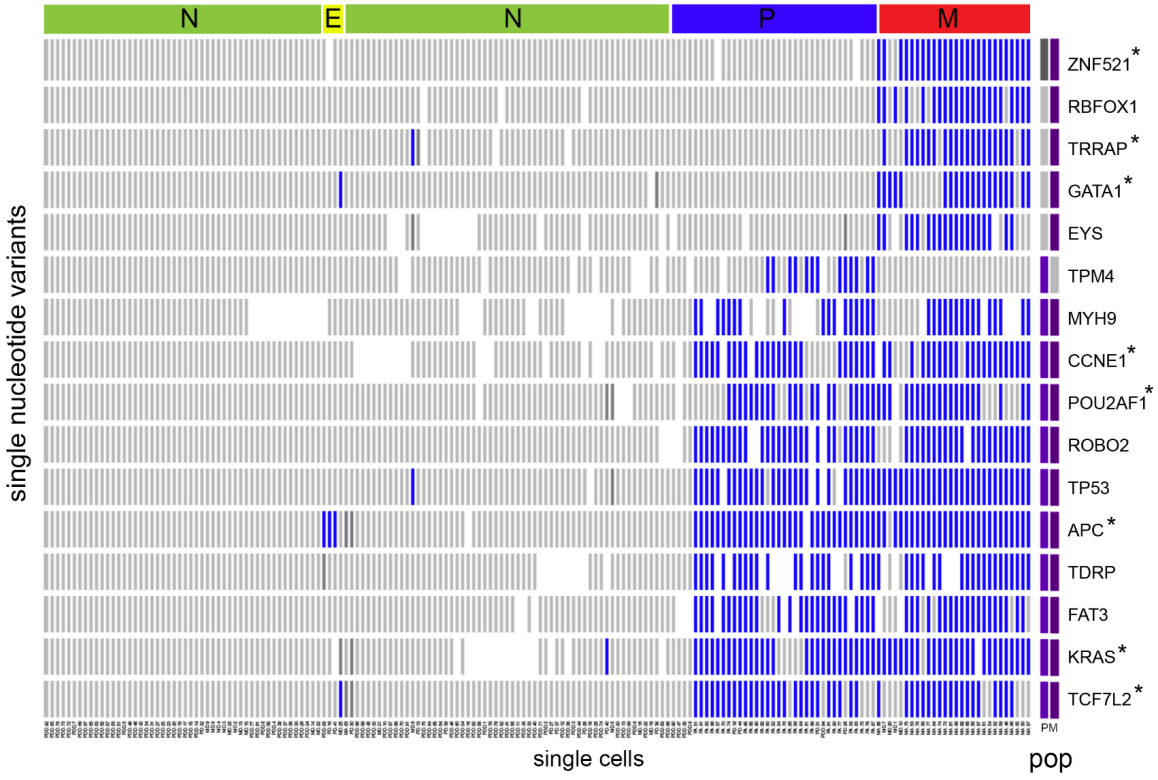

Figure 44 - SNV Mutation Heatmap for CO5

The heatmap is built using two-dimensional hierarchical-clustering. Each row represents a SNV and each column represents a sample. Blue represents the variant allele. Grey represents the reference allele. White represents low coverage. Dark grey represents false positive errors. The asterisks next to the gene names represent that the SNVs are nonsynonymous mutations. The top bar represents the clusters of sample. N represents normal cells. P represents primary cells. M represents metastatic cells and E represents the progenitor cells with APC mutation only.

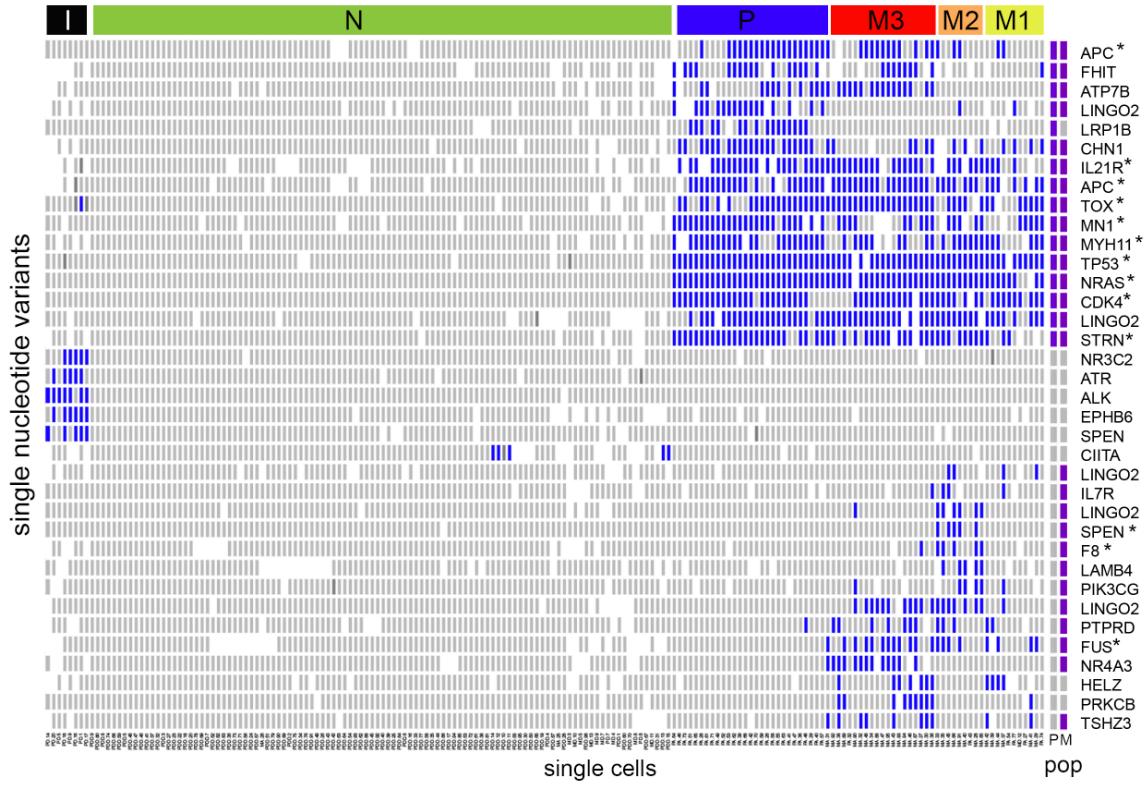

Figure 45 - SNV Mutation Heatmap for CO8

The heatmap is built using two-dimensional hierarchical-clustering. Each row represents a SNV and each column represents a sample. Blue represents the variant allele. Grey represents the reference allele. White represents low coverage. Dark grey represents false positive errors. The asterisks next to the gene names represent that the SNVs are nonsynonymous mutations. The top bar represents the clusters of sample. N represents normal cells. P represents primary cells. M1 – M3 represents metastatic cell subpopulations. I represents the independent lineage of diploid cells.

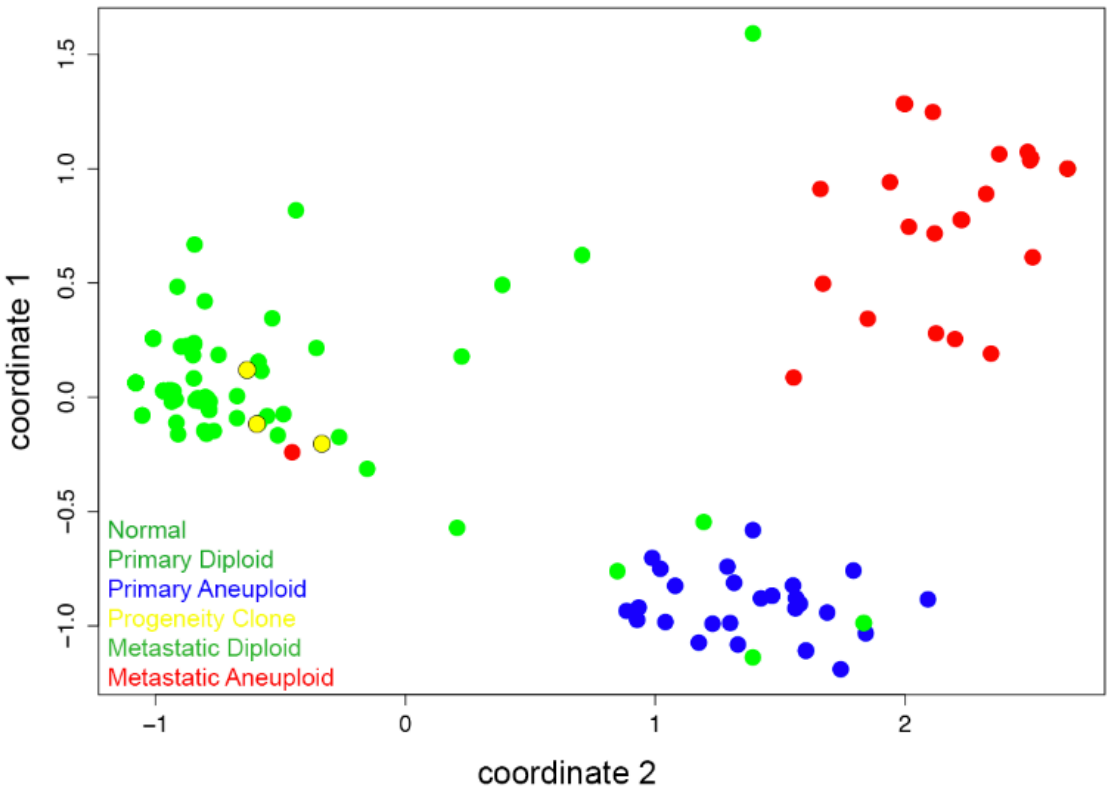

Figure 46 - SNV Multi-Dimensional Scaling Plot for CO5

MDS plot is built using CO5 mutation matrix calculated using Euclidean distance. Green represents normal sample, primary diploid and metastatic diploid cells. Blue represents primary aneuploid samples. Red represents metastatic aneuploid samples. Yellow represents the three progenitor cells with APC-only mutation.

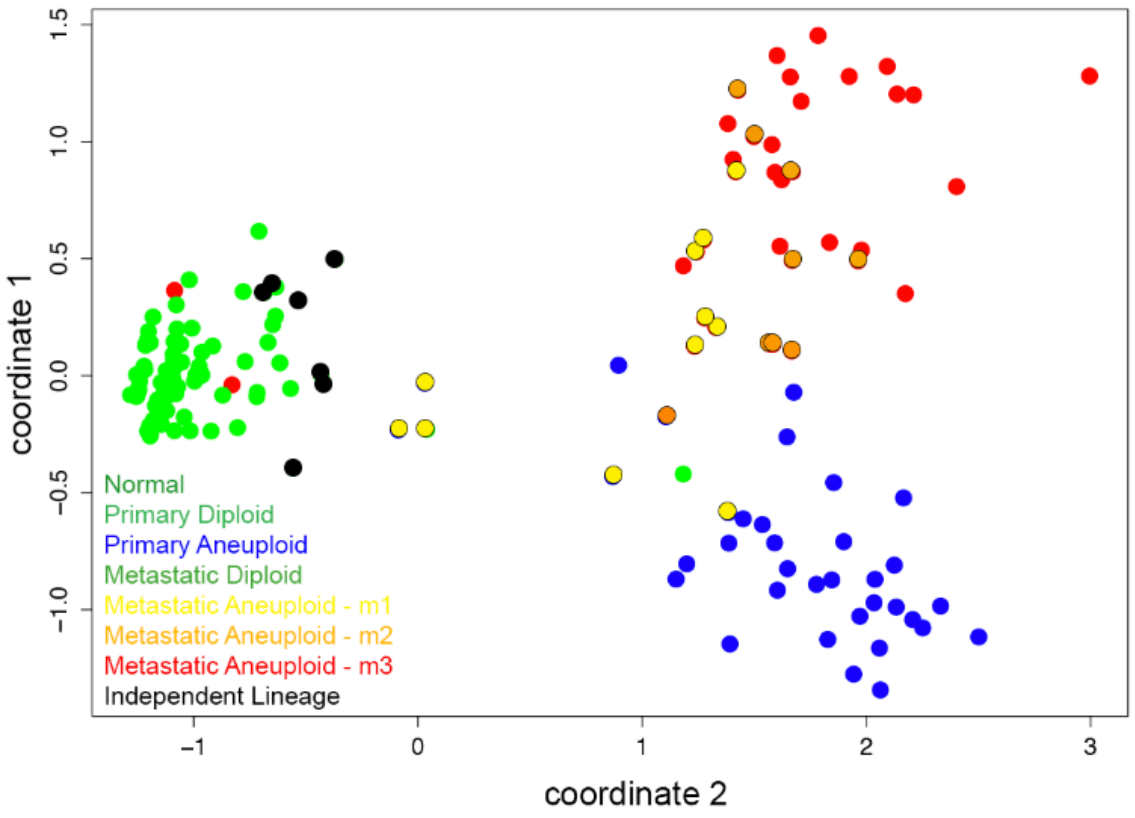

Figure 47 - SNV Multi-Dimensional Scaling Plot for CO8

MDS plot is built using CO8 mutation matrix calculated using Euclidean distance. Green represents normal sample, primary diploid and metastatic diploid cells. Blue represents primary aneuploid samples. Yellow, orange and red represent three metastatic aneuploid cell subpopulations. Black represents the diploid cells sharing the independent lineage.

are three distinct populations (diploid, primary aneuploids and metastatic aneuploids) that clustered by themselves. There are rare aneuploid cells that are clustered in the diploid cells (and same for vice versa). This is caused by the flow cytometer mistakenly flow sorted an aneuploid cell as a diploid cell due to the closeness of two ploidy peaks.

6.2.3 Copy Number Analysis of Single Tumor Cells

In addition to SNVs, we also performed copy number profiling of single cell using single nucleus sequencing in CO5 and CO8.[72] (Figure 36) Using degenerate-oligo-primed PCR, we amplified and sparse-sequenced the genome of single cells. We surveyed 8 primary and 15 metastatic cells in CO5, as well as 19 primary and 24 metastatic cells in CO8. The single cell copy number profile were analyzed using clustered heatmaps, which show the amplifications and deletions across the genomes for both patients. (Figure 48 and 49)

For CO5, the primary and metastatic cells show highly similar profiles with shared focal deletion in the 8p arm and amplification in 7, 13, and 20q arm, but were distinguished by amplification in chromosome X amplification in the metastatic cells. (Figure 48) There is no subpopulation clustered within each primary and metastatic site. For CO8, the primary and metastatic cells each show distinct copy number profiles, with subclonal profiles within the metastatic population. (Figure 49) The primary cells have amplification in chromosome 9p that the metastatic cells don't have. For the two clusters found in the metastatic cells, they share most major CNV events such as amplifications in chromosome

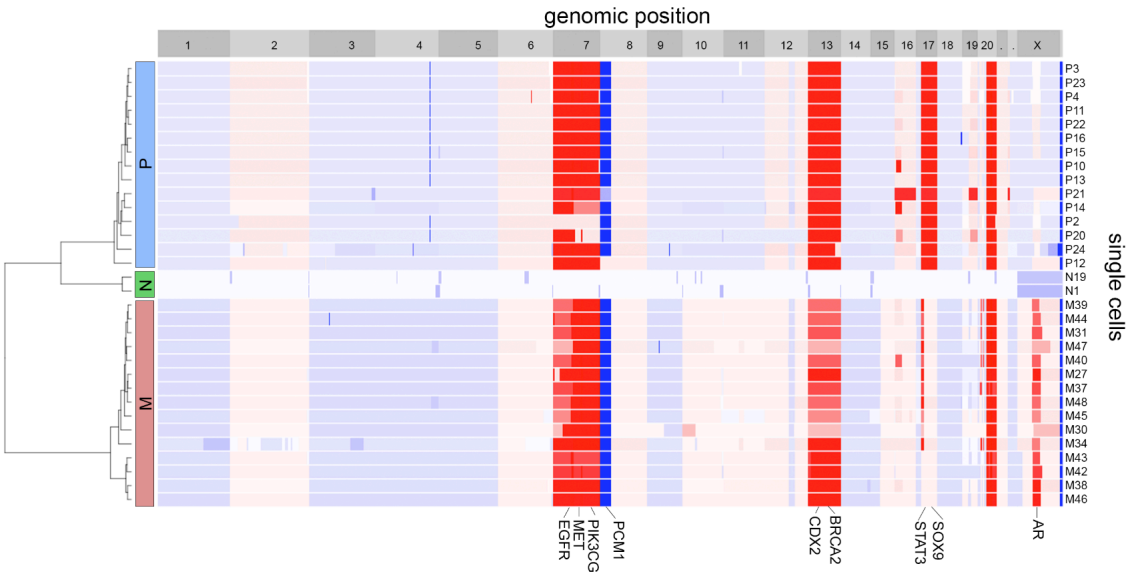

Figure 48 - Copy Number Heatmap for CO5

The heatmap illustrates the copy number status of single cells. Each row represents a single cell. The single cell samples are clustered by hierarchical clustering. Red represents copy number gain and blue represents copy number loss.

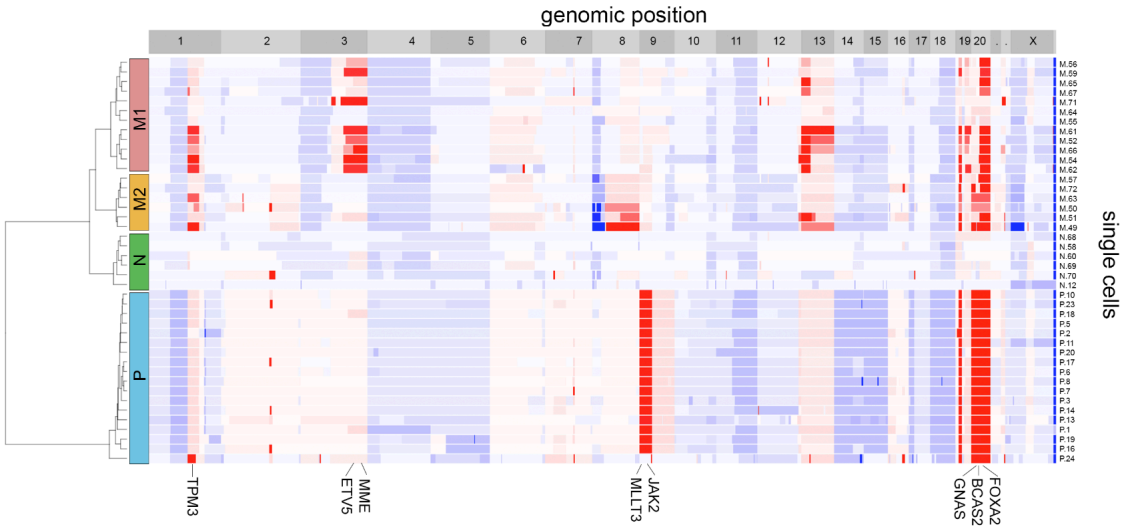

Figure 49 - Copy Number Heatmap for CO8

The heatmap illustrates the copy number status of single cells. Each row represents a single cell. The single cell samples are clustered by hierarchical clustering. Red represents copy number gain and blue represents copy number loss.

1q and 13p. However, cluster 1 (m1) is distinguished by amplification in chromosome 3q.

Next, using the copy number data of CO5 and CO8, we built MDS plots. (Figure 50 and 51) For both patients, MDS plots show three populations (normal samples, primary aneuploid cells and metastatic aneuploid cells) that are distinctly clustered together.

6.2.4 Phylogenetic Analysis

Using the SNVs from CO5 and CO8 as evolutionary markers, we construct phylogenetic trees using maximum parsimony. For CO5, the trees show that the diploid cells are at the base of the tree with the normal samples. (Figure 52 and 53) At the long trunk of the tree, the common mutations (for example, *APC, KRAS, FAT3, MYH9, TDRP, ROBO2, TCF7L2* and *TP53*) are acquired for all the primary and metastatic aneuploid cells. For the metastatic aneuploid cells, additional mutations (*ZNF521*, *RBFOX1*, *EYS*, *TRRAP*, *GATA1*) are then acquired.

For CO8, the diploid cells are at the base of the tree with the normal samples. Mutations (*NRAS,CDK4, TP53, STRN, LINGO, APC, IL21R* and *MYH11*) that are shared in all the aneuploid cells are then acquired at the trunk of tree. (Figure 54 and 55) Then tree is diverged into two main branches. The primary branch has acquired the *LRP1B* mutation while the metastatic branch has acquired *HELZ* and *TSHZ3* mutations. As described in Figure 45, the small cluster of diploid cells is found evolving from the base of the tree, separated from the aneuploid branch. (*NR3C2, ATR, ALK, EPHB6, SPEN* and *CIITA)*

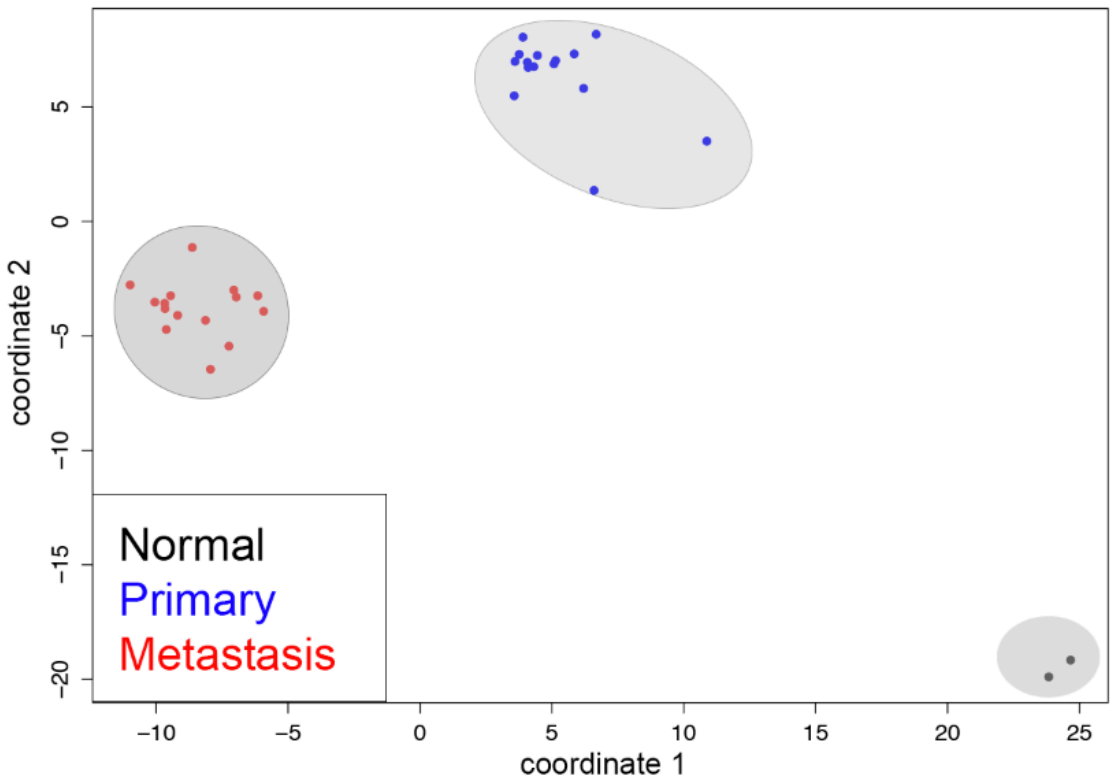

Figure 50 - CNV Multi-Dimensional Scaling Plot for CO5

MDS plot is built using CO5 single cell copy number profiles calculated using Euclidean distance. Black represents normal samples. Blue represents primary aneuploid samples. Red represents metastatic aneuploid samples.

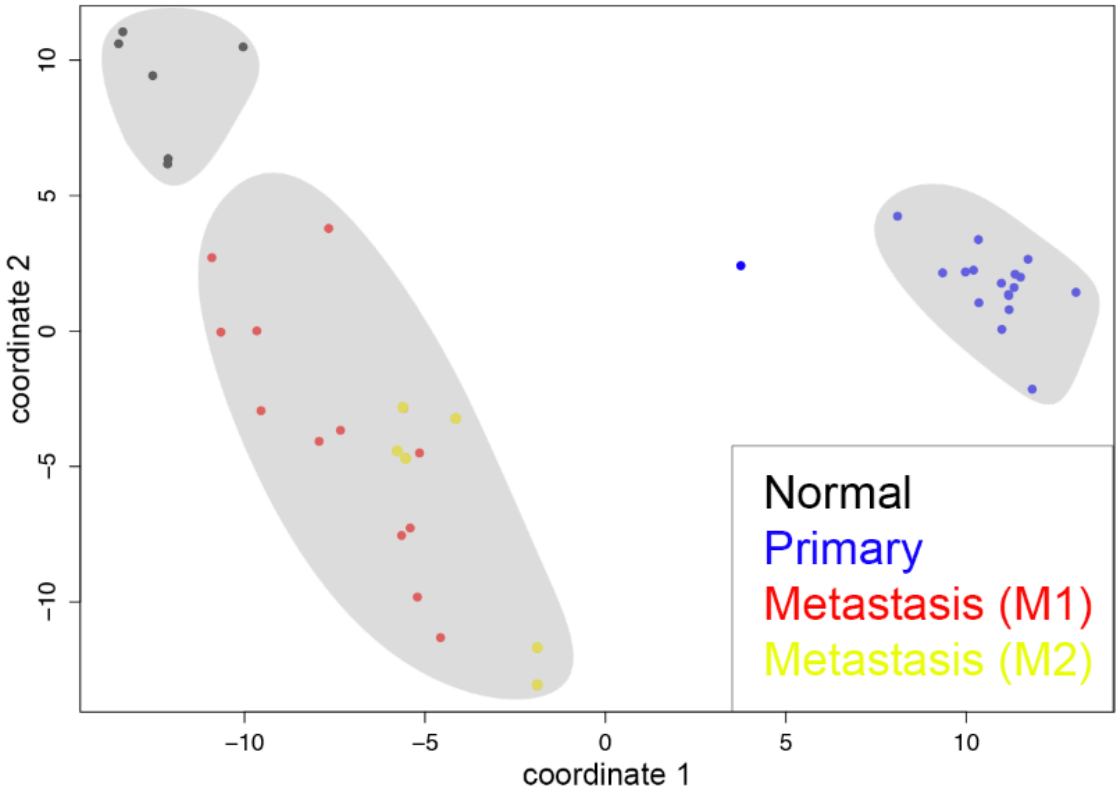

Figure 51 - SNV Multi-Dimensional Scaling for CO8

MDS plot is built using CO8 copy number profiles calculated using Euclidean distance. Black represents normal sample. Blue represents primary aneuploid samples. Red and yellow represent the two metastatic aneuploid subpopulations.

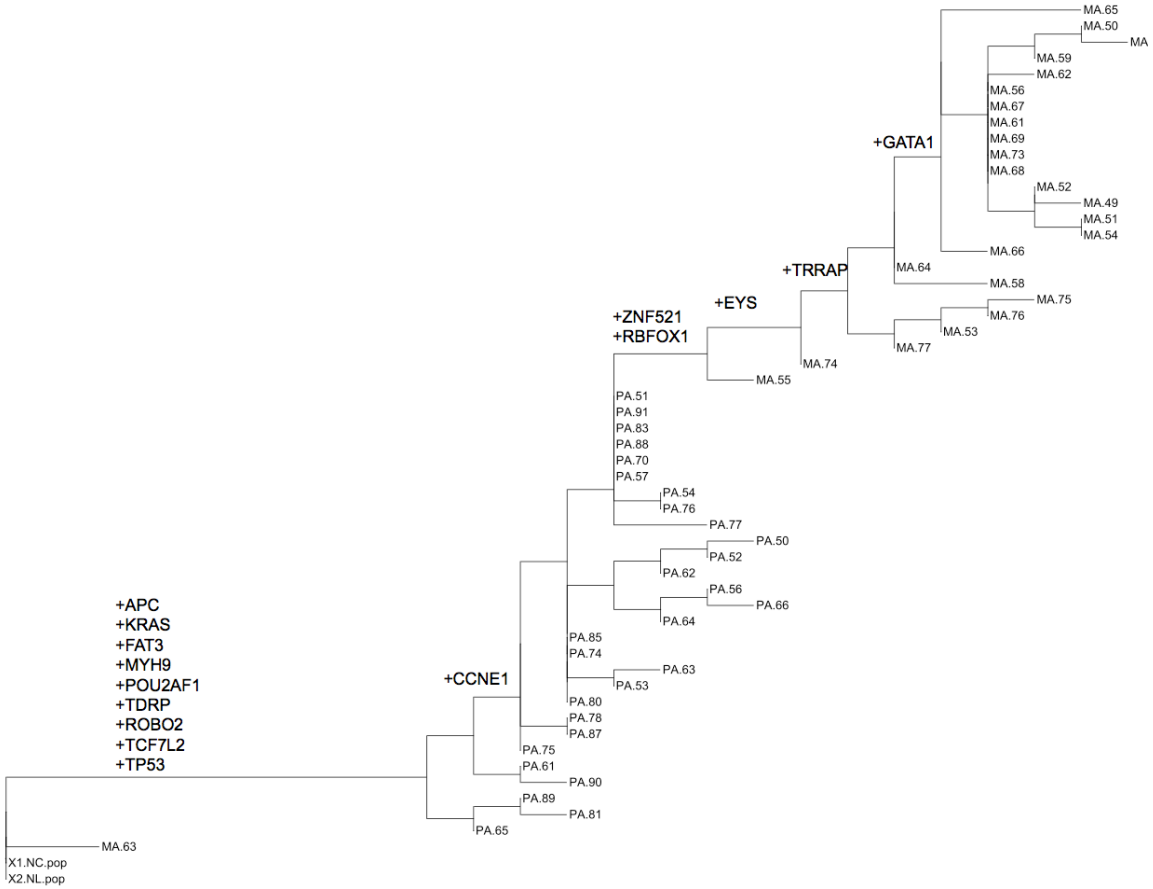

Figure 52 - CO5 Single Cell Phylogenetic Tree with Gene Annotation

The tree is constructed using maximum parsimony using mutation status of each aneuploid cell. The tree is rooted at the node of NC.pop and NL.pop (Normal colon and normal liver tissue)

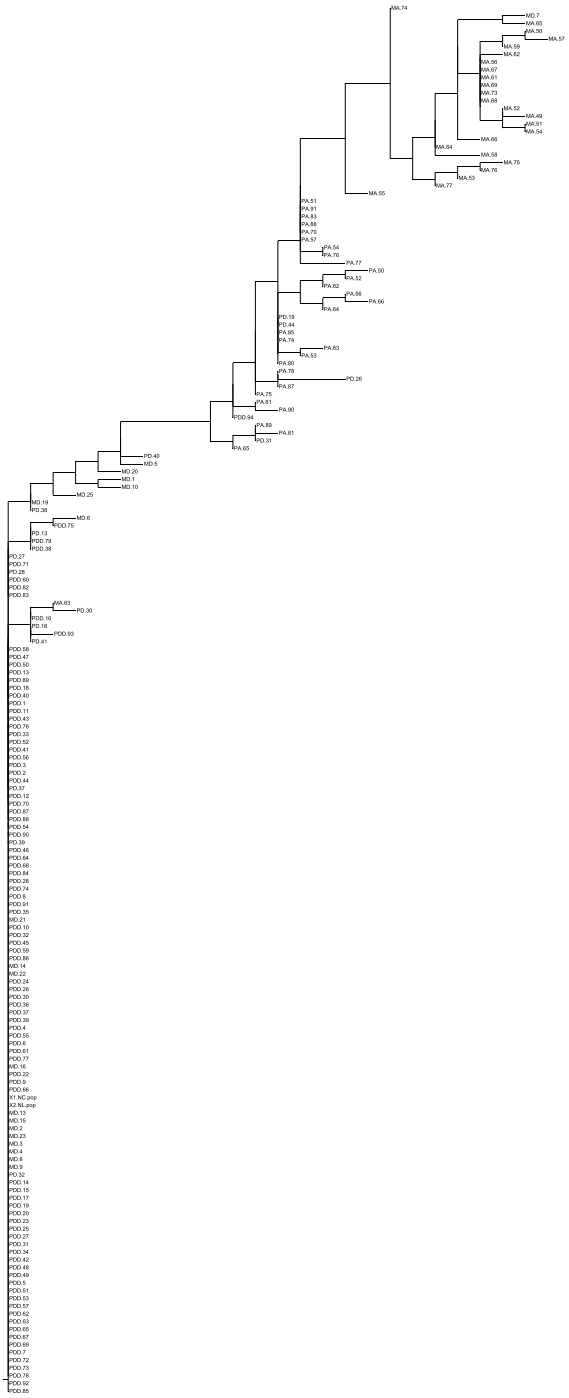

Figure 53 - CO5 Single Cell Phylogenetic Tree with Aneuploid and Diploid Cells

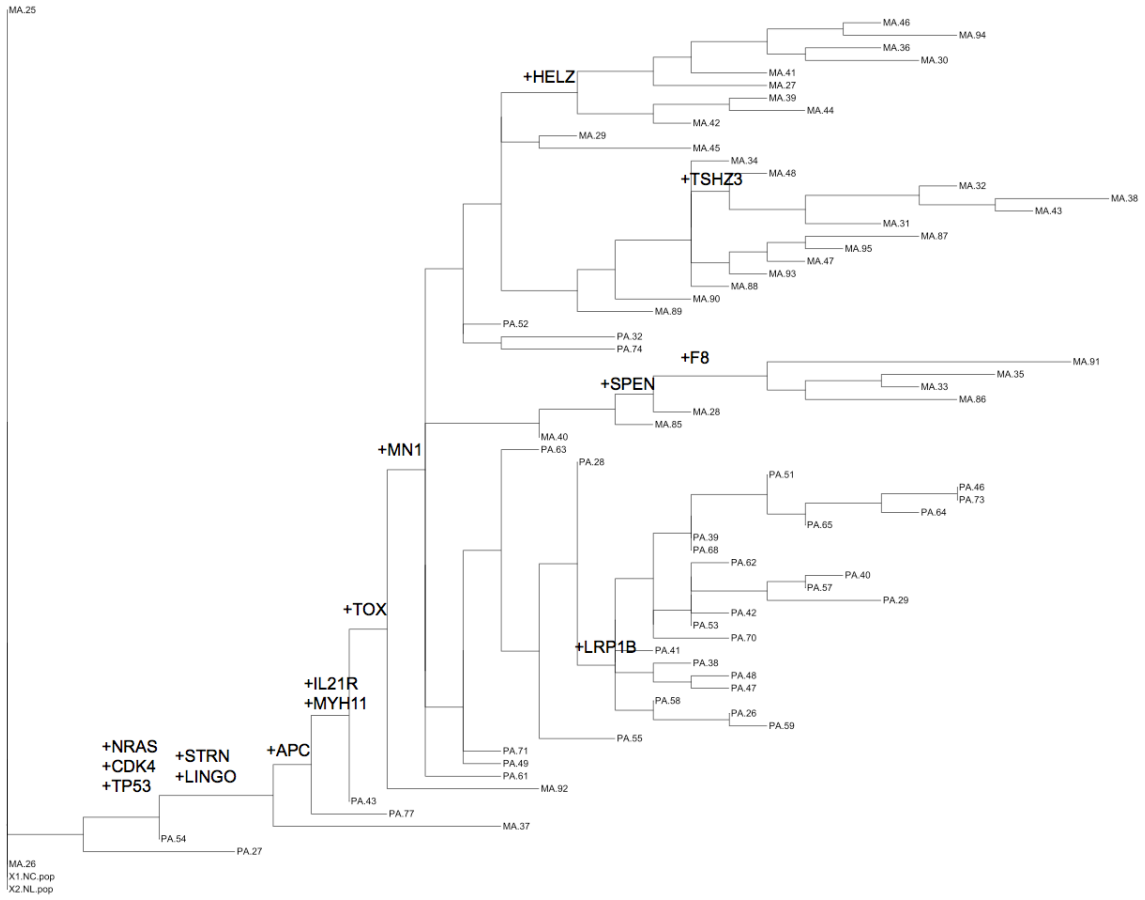

Figure 54 - CO8 Single Cell Phylogenetic Tree with Gene Annotation

The tree is constructed using maximum parsimony using mutation status of each aneuploid cell. The tree is rooted at the node of NC.pop and NL.pop (Normal colon and normal liver tissue)

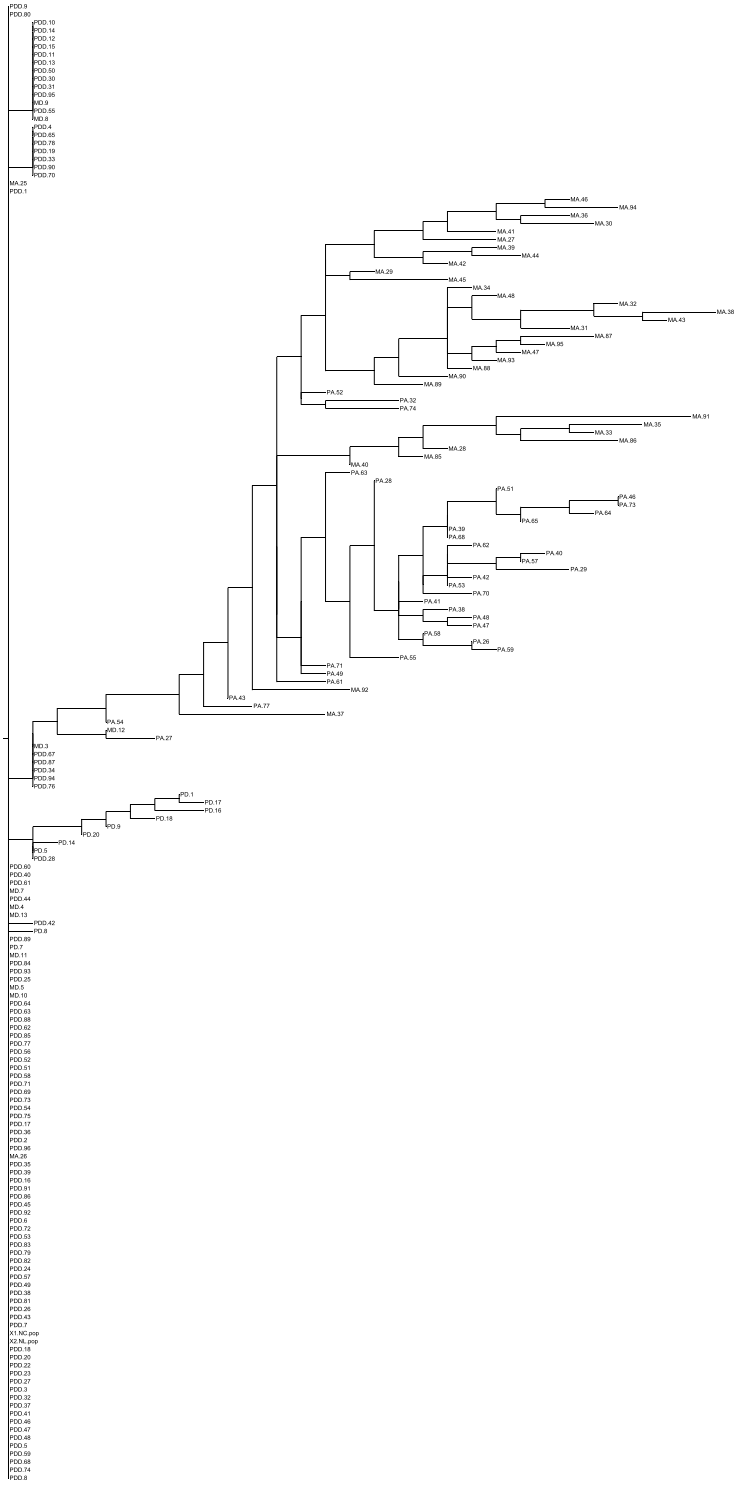

Figure 55 - CO8 Single Cell Phylogenetic Tree with Aneuploid and Diploid Cells

6.3. Discussions

In this chapter, we have described how CRC tumor cells from two patients metastasize using SCS. For CO5, we found each primary and metastatic tumor to be homogeneous. The SNVs and CNVs are highly similar across all cells in each tumor. All single cell data suggests that this patient follows the late dissemination model. For CO8, we found the mutations to be highly similar across all cells in the primary tumors, however there are three subpopulations in the metastasis in both SNVs and CNVs data.

Interestingly, we found three diploid cells that bear only the *APC* mutations in CO5. We suspect that these cells are part of the progenitor clone that gave rise to the overall tumor. In CO8, we found a subclone of diploid cells with 5 mutations. The phylogenetic tree suggests that this subclone of diploid cells has a separate lineage than those of aneuploid cells because this diploid subclone has no shared mutations with the overall tumor.

The detection of the three subclones within the CO8 metastatic tumor is important to note because it highlights the potentials of SCS in capturing the intratumor heterogeneity. By flow-sorting nuclei from the overall tumor, we could investigate the intratumor heterogeneity by broadly sampling single cells without bias. The approach allowed us to obtain single cell genomic data that is representative of the overall tumor. This is in contrast to spatial sequencing (as described in **1.4.2 NGS and Intratumor Heterogeneity**), whereby a tumor is cut into multiple sections and sequenced. The sections cut for spatial sequencing

may not be proportional of the subclones in the tumor, whereas random sampling of single cells is representative of the subclone distribution.

In summary, we have demonstrated the metastatic lineage and intratumor heterogeneity in two CRC patients using SCS. To our knowledge, no prior studies have used SCS to study how CRC tumors metastasize. Furthermore, this data would be difficult to replicate using conventional sequencing methods.

CHAPTER SEVEN - DISCUSSION, CONCLUSIONS AND FUTURE DIRECTIONS

Chapter 7 – Discussion, Conclusions and Future Directions

7.1 Discussion and Conclusions

7.1.1 Using Single Nuclei as Input Materials for Sequencing

In this dissertation, we describe our efforts to develop single cell DNA sequencing methods, and their application to study how CRC tumors evolve by tracing the metastatic lineage. Although we have gone through multiple stages of SCS development, we continue to prefer using nuclei as input material. There are multiple reasons why we prefer to use nuclei instead of intact cells.

1. Although it is common to isolate intact single cells from fresh tissues using enzymatic or physical dissociation, it is difficult to achieve the single cell isolation in frozen tissues because cell membranes are ruptured during the freeze-thaw cycles, while nuclear membranes remain intact.
2. Instead of using live/dead cell viability dye, nuclei can easily be stained for ploidy distribution using DAPI, allowing the selection of aneuploid tumor cells. Furthermore, DAPI can be used to avoid the collection of replicating cells in S-phase or the genomes of highly degrade cells (<1n).
3. In highly connected tissues, like neuronal cells, it is difficult to dissociate cells, thus two or more cells may be missorted into one reaction (doublets) and treated as one cell. Nuclei on the other hand can be more easily separated and deposited accurately for flow-sorting applications.

7.1.2 Sequencing Single Cells at the G2/M Phase and at the Aneuploid Peak

The singular challenge of SCS is the low-input material. In order to construct libraries, single cell genome (6 pg) must be amplified more than ten-thousand-fold using WGA polymerases. However, the amplification process is not perfect, thus introducing technical errors (false positive) in the amplified products. Moreover, because there are only two alleles in diploid cells, one allele may get amplified preferably over the other allele, leading to allelic dropout. We address these issues by sequencing cells specifically in the G2/M phase. During the G2/M phase of the cell cycle, the DNA content doubles. Instead of one copy of each allele, there are two copies, thus increasing the chance of polymerase detecting both alleles for amplification.

As described in this dissertation, we compare the sequencing metrics and error rates of cells in G1/0 and G2/M phases. We found that G2/M cells constantly outperformed G1/0 cells in coverage uniformity and coverage breadth (96%). More importantly, we demonstrated that G2/M cells have lower allelic dropout rate (21%) and higher detection efficiency for SNVs (92.37%).

In Chapter 6, we demonstrated using our developed methods on aneuploid cells to study tumor evolution. By selecting aneuploid cells, it ensures that we are sequencing tumor cells, instead of stromal cells, fibroblast or immune cells. Moreover, aneuploid cells have more than two alleles in at least some parts of the genome, decreasing the allelic dropout events that occur during the WGA step.

However, there are caveats and biases to consider when selecting aneuploid cell or G2/M cells for sequencing.

1. In tissues where cells are proliferating slowly, it is difficult to detect a G2/M peak to gate during flow sorting.
2. In tumors where aneuploid cells are present, it is important to also consider the diploid cells. There may be tumor cells present in the diploid peak that have not acquired aneuploidy.
3. When sequencing G2/M tumor cells, fast-proliferating cells are selected and slow-proliferating cells that may have different genomic profiles may be missed (for example, cancer stem cells).
4. Although G2/M cells and aneuploid cells have significantly lower allelic dropout rate (Figure 27), ADO errors should still be considered when interpreting single cell data. For data with ADO rate of 20%, there are 10% of AB alleles dropped to AA, thus 10% of the variant alleles are missed. This is important to consider when interpreting the mutational heatmaps (Figure 44 and 45), whereby the variants are not detected in few aneuploid cells (such as CCNE1 and ROBO2 in CO5 heatmap in Figure 44).

7.1.3 Determining the Number of Single Cells Required for Sampling

As the sequencing throughput of NGS technologies continues to improve and the cost per sample continues decreases, it will become more affordable to sequence thousands of single cells to study intratumor heterogeneity in tumors. In Chapter 6, we sequenced 186 single cells from each patient because we were

able to multiplex up to 96 samples in one sequencing lane. However, there is no consensus on the total number of cells that is needed to be sequenced from a tumor. Previous studies have estimated the numbers of cells needed to be sequenced based on the amount of coverage depth required to detect mutations or copy number from each single cell dataset, as shown in Figure 56.[91] In order to sequence more cells, previous studies increased the cell numbers and compensated by decreasing the total number of bases covered in each cell, hence reducing the sequencing output requirement for each cell. This approach allows researchers to logistically determine the maximum number of single cells that can be sequenced based on the funds available for the project.

However, this approach determines the sampling numbers based on technical needs, and does not address how many cells needed to be sampled to resolve intratumor heterogeneity. The number of single cell needed to be sequenced can be determined using statistical method, as shown as below in which the power of detection determined for sampling number based on the sensitivity required to detect subclones.

$$P(d) = 1 - (1 - s)^n$$

In this formula, s represents the subclonal frequency and n represent the number of single cell.[92] Using this approach, we determine the sensitivity of which intratumor heterogeneity can be detected at high confidence. For example, to detect a subclone of 0.01 frequencies at 95% detection power, 300 cells are needed. Similarly, to detect a subclone of 0.1 at 95% detection power, only 30 cells are needed.

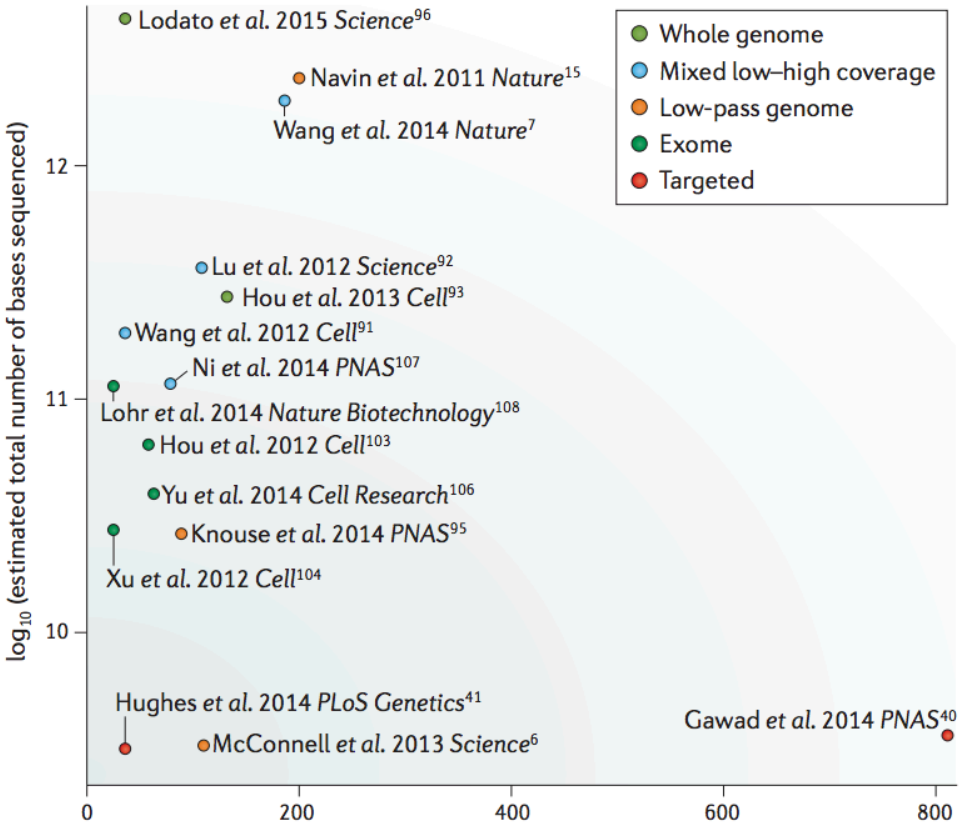

Figure 56 - Number of Single Cell Sequencing Based on Genome Coverage

This figure demonstrates the numbers of single cells past studies have sequenced, as well as the type of genome coverage they used. For studies that sequenced whole genomes, the numbers of cells are below 50. For studies that used sparse sequencing or targeted sequencing, the number of cells is comparatively higher.

(Modified and reproduced from Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature reviews. Genetics* **17**, 175-188, doi:10.1038/nrg.2015.16 (2016). with permission from Nature Publishing Group)

7.1.4 Single Cell Sequencing Applications in Clinical Diagnostics

In Chapter 6, we used SCS to understand metastatic lineage in CRC, and found that both patient followed a late-dissemination model. We also found three single primary cells in CO5 that had the *APC* mutation, and we suspect this cell might be a part of the progenitor clone that gave rise to the overall tumor. Moreover, we discovered three subclonal populations within the metastatic tumor in CO8. For both patients, we used single cell data to construct phylogenetic trees to trace the metastatic evolution. These findings would be difficult to discover with conventional NGS methods due to low sensitivity and admixtures of disparate genomes in the bulk tissues. This dissertation has demonstrated that SCS is a powerful tool for understanding the fundamental basis of metastasis in colorectal cancer. Similarly, colorectal cancer is a suitable model to demonstrate the advantage of SCS because the primary colon tumors are often excised along with liver hepatectomy before treatments are administered. Treatment-naïve tumor samples are important because SCS data would allow us to understand the innate evolution of the tumor, instead of the genomic state of chemo-resistant tumors under selective pressure.

The advent of SCS also provides an invaluable tool for clinical research and diagnostics. In non-invasive monitoring of the blood, SCS can be used to profile circulating tumor cells (CTCs). CTCs serve as a link between the primary tumor and metastasis. However, it has been difficult to investigate the genomes of CTCs due to the rare frequency in the blood (1-10 CTCs out of millions of leukocytes in 1 mL of blood in patients with metastatic diseases).[93] Previous

studies have used copy number profiling to study CTCs of lung cancer patients.[94] However this approach cannot identify mutation at base pair resolution. A recent study attempted to exome-sequence 19 CTCs from a prostate cancer patient.[95] They found that mutations in CTCs are concordant to those in the metastatic tumor. However, these mutations were detected by combining all single cell libraries; this is because of low coverage breadth and high error rates of their SCS method. This approach is similar bulk sequencing, thus nullifying the intention of SCS.

In addition to the cancer field, SCS has provided tremendous improvements in the reproductive field. To prevent miscarriage and genetic disorder due to aneuploidy or mutations, previous study has shown to use SCS to screen and preselect fertilized egg.[96] Sequencing the oocyte polar bodies can detect genetic disorder from the mother. Similarly, sequencing cells from the blastocyst stage of the embryo can detect genetic disorder from the father.[96]

However, there is a major challenge in utilizing SCS in clinical settings. NGS tests are often validated by another method, such as Sanger Sequencing. For SCS, validation using the original single cell DNA is impossible because all DNA from the particular single cell has been used to amplify. Amplified single cell DNA is not suitable for validation because technical errors have already been introduced during the amplification step, hence incorrectly categorized as true mutation. An indirect approach to circumvent this issue is to annotate mutations that occur in two or more cells. As we have shown in Chapter 4, our SCS methods have false positive error rate of approximate 3.2e-5.[88] By detecting

mutations in two or more cells, false positive error rate would decrease to 1.02e-9, which is sufficient to confidently call mutations. However, this approach would miss the *de novo* mutations present in only one cell.

7.1.5 Late Dissemination Model in Colorectal Cancer

The process of metastasis was likened to a decathlon in a previously published review article by James Talmadge and Isaiah Fidler, whereby a cancer cell must overcome multiple steps of obstacles in order to successfully metastasize.[97] These steps include the vascularization of tumor mass, the local invasion of the host stroma by tumor cells, the detachment of tumor cells into circulation, the survival of tumor cells that trafficked through the circulation and arrest in a capillary bed, the extravasation and proliferation of tumor cells at distant organ sites and the re-establishment of vascularization in the metastasis.[97] However, it remains unclear what mutations are needed to acquire in order to overcome these steps of obstacles. Moreover, this review article had also stated that the time required to form primary tumor and develop metastasis was different for patients with different cancer types. For example, mammography has shown that the growth of breast tumors require an average of 12 years to grow from initiation to a size of 1 cm, whereas it takes up to 25 years for familial adenomatous polyposis to become malignant. This may be caused by the fact that different cancer types may follow different tumor progression models and different models of metastasis. (Figure 1 and 2)

Our single cell sequencing data identified a late-dissemination model in two mCRC patients. In the previous studies, different models of metastasis are

difficult to distinguish using conventional NGS methods. (Figure 1) The late-dissemination model posits that the primary tumor develops for a long period of time before tumor cells disseminate and seed at distant organs. In this model, the majority of mutations are acquired and accumulated in the primary tumor, and the cells carrying these mutations metastasize but do not acquire many additional mutations after seeding. Another model, self-seeding, posits that tumor cells travel bi-directionally between the primary and metastatic sites; this leads to primary and metastatic sites sharing the majority of mutations, similar to the late-dissemination model. Conventional NGS methods have difficulty distinguishing between these two models because both suggest that the majority of mutations are shared at both sites. In contrary, SCS can distinguish between these models by detecting mutations specific to each single cell because it can detect individual tumor cells from the metastatic tumors that have re-seeded the primary tumor.

Our data, as well as previous CRC studies, support a late-dissemination model in CRC without any evidence for self-seeding or early dissemination. This model has major implications for CRC patients by suggesting that surgical resection or therapeutic treatment of the primary CRC tumor can prevent metastasis. Furthermore, the late-dissemination model suggests that the primary and metastatic tumor shared similar genomic profile. A tumor biopsy should represent other tumor sites within a patient, thus a single tumor biopsy is enough to guide therapy for metastatic patients. As it was suggested that polyps require more than 10 years to become malignant, early detection of polyps in the colon

(with the use of colonoscopy) is beneficial to the management and treatment of early-stage primary tumor. Removing the polyps or early-stage adenoma in the colon would prevent metastasis to develop, thus decreasing the prevalence of stage IV CRC.

7.1.6 Identification of Progenitor Clones Using Single Cell Sequencing

In addition to the analysis of aneuploid tumor cells, we investigated the diploid cells flow-sorted from the two CRC patients. While most of the diploid cells did not have any mutations present, we did identify three diploids cells in CO5 that showed only a single mutation in a cancer gene: *APC*. This *APC* mutation (c.C4012T) was present in every tumor cell sequenced from the primary and metastatic liver tumors, suggesting that it was acquired at the earliest stages of tumor initiation. This finding agrees with the previously described Vogelgram, in which described the stages of CRC genetic alterations and the *APC* mutation is the first event of CRC tumorigenesis.[98] (Figure 57)

The detection of this progenitor clone is important because it has been challenging to observe in previous studies. When a given tumor is sequenced, only the mutations present at the time when tumor is surgically excised is detected. The original cell with the first 'hit' mutation, or the progenitor cell, is not detected because it is outcompeted by other clones with subsequent additional mutations that have increased fitness. We were able to detect these three cells because they have not been completely outgrown by the major subpopulations.

This finding highlights the advantage of using SCS to study tumor evolution and metastasis. Out of 112 primary diploids that were sequenced, we

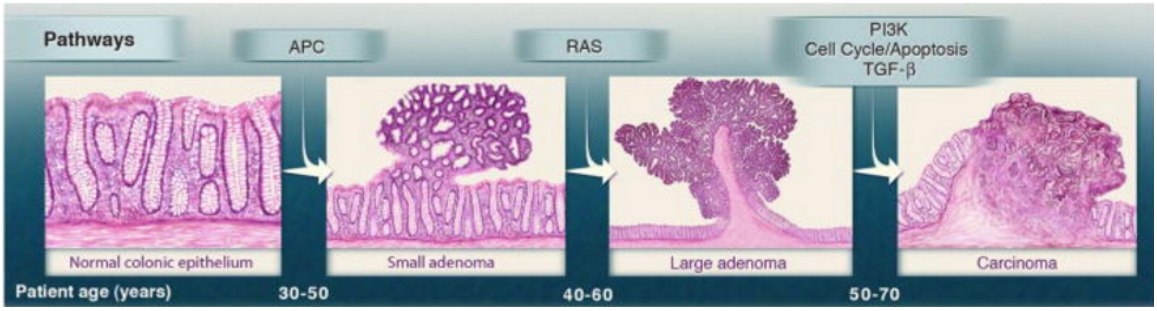

Figure 57 - Vogelgram describing the Events of Genetic Alterations in CRC

The sequential mutations occurred during CRC has been previously described by Bert Vogelstein. The first 'hit' is usually the *APC* mutations, leading the activation of oncogenes, such as *KRAS*, and inactivation tumor suppressor genes, such as *TP53*.

(Modified and reproduced from Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013) Permission is obtained from the publisher, The American Association for Advancement of Science)

found only 3 cells with the *APC* mutations; this is equivalent to the frequency of 0.0268. The low frequency would be difficult to detect using conventional sequencing without high coverage depth.

7.2 Future Directions

While SCS shows great promise in cancer biology as well as other fields of biology, further technical improvements are needed before it can become widely used in research and clinical settings. Because SCS is a complex method encompassing multiple techniques (single cell isolation, whole-genome amplification, library preparation, targeted capture and data analysis), fine-tuning of each of these techniques can improve the overall performance of SCS. For example, single cells isolation can be more accurately and precisely performed with microfluidics and robotics. Improving DNA polymerase can decrease the technical errors generated during single cell genome amplification. Novel bioinformatics software may distinguish the real biological variants from technical errors during data analysis. These are the areas that we foresee improving in coming years in order to optimize the ability of SCS to address novel biological questions and solve current clinical challenges.

To improve the prognosis of patients with metastatic CRC, we believe that SCS of CTCs is one of the important future steps to fully understand the metastatic cascade. As described in **7.1.4**, CTCs are key intermediates of metastasis. Past studies have used the enumeration of CTCs as a prognostic marker for breast cancer patients.[99] It was found that patients with 5 or more CTCs in 7.5mL of blood had a higher chance of relapse after treatments and

metastasis, when compared to those with less than 5 CTCs.[99] However, it remains unclear how the numbers of CTCs correlate to patient survival. It would be interesting to use SCS to discern the genomes of CTCs from these two groups of patients. In other words, SCS may give us insight to understand whether certain mutations in CTCs increase the metastatic capabilities and potentially resist treatments.

Furthermore, we predict SCS to be widely used as a non-invasive monitoring tool of mutation status of CRC tumor genome, and aid in determining the appropriate treatments. For example, it was recently shown that immunotherapy targeting the programmed-death 1 (PD-1) pathway using pembrolizumab is effective in MSI-high CRC patients.[100] However, the MSI status of CRC patients is not unknown without performing tumor biopsy. SCS can be an alternative approach (liquid biopsy) to tumor biopsy, where MSI status is determined the number of mutations in CRC CTCs detected by SCS, and oncologists can decide whether immunotherapy is needed to administer. Moreover, there are cancers in specific organs that are difficult to safely biopsy. For example, there are 15% of patients who suffers partial lung collapse when undergoing lung needle biopsy, and there are 1% of patients resulting in excessive bleeding.[101] In contrary, liquid biopsy of CTC using SCS is a much safer approach.

Nevertheless, CTCs remain difficult to study due to the rarity of these cells and the technical flaws of SCS. Technical improvement in decreasing SCS error rates will allow researchers to answer the following opposing questions:

1. Do CTCs possess the mutations of primary tumors and later acquire additional metastatic-specific mutations after seeding?
2. Do CTCs have already acquired mutations and possess the metastatic capability necessary to proliferate in distant organs?

Additionally, we also anticipate future studies to focus on both the genomic and transcriptomic profiles of CTCs. This can be achieved by developing novel methods in the sequencing genome and transcriptome of the same cell. A method was recently published that can detect DNA copy number and gene expression of one cell.[102] However, single nucleotide variants (SNVs) cannot be detected due to the problems with low coverage. Further refinements of the method by increasing the coverage will allow researchers to understand how changes in single nucleotide variants can lead to changes in gene expression in individual cells. For CTCs, this proposed technique would allow us to investigate how the combination of mutations and epithelial-to-mesenchymal transition (EMT) expressions in CTCs can potentially affect the metastatic potential.

In Chapter 6, we detected three cells that harbored only the APC mutation, suggesting that they were a part of the progenitor clone that gave rise to the overall tumor mass. The progenitor clone has been difficult to detect in previous genomic studies because it is usually outcompeted by the more proliferative clones. To further understand the sequential event during CRC tumorigenesis, as shown in the Vogelgram in Figure 57, we can perform SCS on a patient with lesions of multiple stages (polyps, adenoma, carcinoma). By

sequencing single cells on the lesions of multiple stages in the same patient, we can track when a specific mutation occurs and how different clones expand and shrink during tumorigenesis.

In summary, we expect that the use of SCS will become widely utilized, not only in cancer biology, but also in other aspect of biology, such as microbiology, developmental biology and prenatal genetic diagnosis, in which SCS can provide unprecedented genomic information and improve our understanding of variety of human pathologies.

References

1 American Cancer Society. *What are the key statistics about colorectal cancer?*, <http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-key-statistics> (2015).

2 Howlader N, N. A., Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z,Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA. *SEER Cancer Statistics Review, 1975 - 2011*, 2013).

3 American Cancer Society. Colorectal Cancer Facts & Figures 2011-2013. (2011).

4 Kelloff, G. J., Schilsky, R. L., Alberts, D. S., Day, R. W., Guyton, K. Z., Pearce, H. L., Peck, J. C., Phillips, R. & Sigman, C. C. Colorectal adenomas: a prototype for the use of surrogate end points in the development of cancer prevention drugs. *Clin Cancer Res* **10**, 3908-3918, doi:10.1158/1078-0432.CCR-03-0789 (2004).

5 American Cancer Society. *American Cancer Society Guidelines for the Early Detection of Cancer*, <http://www.cancer.org/healthy/findcancerearly/cancerscreeningguidelines/american-cancer-society-guidelines-for-the-early-detection-of-cancer> (2015).

6 Edwards, B. K., Ward, E., Kohler, B. A., Eheman, C., Zauber, A. G., Anderson, R. N., Jemal, A., Schymura, M. J., Lansdorp-Vogelaar, I., Seeff, L. C., van Ballegooijen, M., Goede, S. L. & Ries, L. A. Annual report to the

nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer* **116**, 544-573, doi:10.1002/cncr.24760 (2010).

7 Baxter, N. N., Goldwasser, M. A., Paszat, L. F., Saskin, R., Urbach, D. R. & Rabeneck, L. Association of colonoscopy and death from colorectal cancer. *Annals of internal medicine* **150**, 1-8 (2009).

8 Wong, S. H., Sung, J. J., Chan, F. K., To, K. F., Ng, S. S., Wang, X. J., Yu, J. & Wu, W. K. Genome-wide association and sequencing studies on colorectal cancer. *Seminars in cancer biology* **23**, 502-511, doi:10.1016/j.semcancer.2013.09.005 (2013).

9 Pritchard, C. C. & Grady, W. M. Colorectal cancer molecular biology moves into clinical practice. *Gut* **60**, 116-129, doi:10.1136/gut.2009.206250 (2011).

10 Trautmann, K., Terdiman, J. P., French, A. J., Roydasgupta, R., Sein, N., Kakar, S., Fridlyand, J., Snijders, A. M., Albertson, D. G., Thibodeau, S. N. & Waldman, F. M. Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clin Cancer Res* **12**, 6379-6385, doi:10.1158/1078-0432.CCR-06-1248 (2006).

11 Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

12 American Cancer Society. *How is colorectal cancer staged?*, <http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-staged> (2015).

13 American Cancer Society. *Treatment of colon cancer by stage*, <http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-treating-by-stage-colon> (2015).

14 American Cancer Society. *What are the survival rates for colorectal cancer by stage?*, <http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-survival-rates> (2015).

15 Mehlen, P. & Puisieux, A. Metastasis: a question of life or death. *Nat Rev Cancer* **6**, 449-458, doi:10.1038/nrc1886 (2006).

16 Klein, C. A. Parallel progression of primary tumours and metastases. *Nat Rev Cancer* **9**, 302-312, doi:10.1038/nrc2627 (2009).

17 Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197-200 (1975).

18 Sosa, M. S., Bragado, P. & Aguirre-Ghiso, J. A. Mechanisms of disseminated cancer cell dormancy: an awakening field. *Nat Rev Cancer* **14**, 611-622, doi:10.1038/nrc3793 (2014).

19 Paez, D., Labonte, M. J., Bohanes, P., Zhang, W., Benhanim, L., Ning, Y., Wakatsuki, T., Loupakis, F. & Lenz, H. J. Cancer dormancy: a model of early dissemination and late cancer recurrence. *Clin Cancer Res* **18**, 645-653, doi:10.1158/1078-0432.CCR-11-2186 (2012).

20 American Cancer Society. *What is a cancer of unknown primary?*, <http://www.cancer.org/cancer/cancerofunknownprimary/detailedguide/cancer-unknown-primary-cancer-of-unknown-primary> (2014).

21 Kim, M. Y., Oskarsson, T., Acharyya, S., Nguyen, D. X., Zhang, X. H., Norton, L. & Massague, J. Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315-1326, doi:10.1016/j.cell.2009.11.025 (2009).

22 Comen, E., Norton, L. & Massague, J. Clinical implications of cancer self-seeding. *Nat Rev Clin Oncol* **8**, 369-377, doi:10.1038/nrclinonc.2011.64 (2011).

23 Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A. & Stratton, M. R. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A* **105**, 13081-13086, doi:10.1073/pnas.0801523105 (2008).

24 Gonzalez-Garcia, I., Sole, R. V. & Costa, J. Metapopulation dynamics and spatial heterogeneity in cancer. *Proc Natl Acad Sci U S A* **99**, 13085-13089, doi:10.1073/pnas.202139299 (2002).

25 Macintosh, C. A., Stower, M., Reid, N. & Maitland, N. J. Precise microdissection of human prostate cancers reveals genotypic heterogeneity. *Cancer research* **58**, 23-28 (1998).

26 Shipitsin, M., Campbell, L. L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M., Halushka, M. K., Sukumar, S., Parker, L. M., Anderson, K. S., Harris,

L. N., Garber, J. E., Richardson, A. L., Schnitt, S. J., Nikolsky, Y., Gelman, R. S. & Polyak, K. Molecular definition of breast tumor heterogeneity. *Cancer cell* **11**, 259-273, doi:10.1016/j.ccr.2007.01.013 (2007).

27 Michor, F. & Polyak, K. The origins and implications of intratumor heterogeneity. *Cancer Prev Res (Phila)* **3**, 1361-1364, doi:10.1158/1940-6207.CAPR-10-0234 (2010).

28 Navin, N. E. & Hicks, J. Tracing the tumor lineage. *Molecular oncology* **4**, 267-283, doi:10.1016/j.molonc.2010.04.010 (2010).

29 Brown, T. M. & Fee, E. Rudolf Carl Virchow: medical scientist, social reformer, role model. *American journal of public health* **96**, 2104-2105, doi:10.2105/AJPH.2005.078436 (2006).

30 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

31 Clarke, M. F., Dick, J. E., Dirks, P. B., Eaves, C. J., Jamieson, C. H., Jones, D. L., Visvader, J., Weissman, I. L. & Wahl, G. M. Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer research* **66**, 9339-9344, doi:10.1158/0008-5472.CAN-06-3126 (2006).

32 Mardis, E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry* **6**, 287-303, doi:10.1146/annurev-anchem-062012-092628 (2013).

33 Devilee, P., van Vliet, M., Kuipers-Dijkshoorn, N., Pearson, P. L. & Cornelisse, C. J. Somatic genetic changes on chromosome 18 in breast carcinomas: is the DCC gene involved? *Oncogene* **6**, 311-315 (1991).

34 Bodmer, W. F., Bailey, C. J., Bodmer, J., Bussey, H. J., Ellis, A., Gorman, P., Lucibello, F. C., Murday, V. A., Rider, S. H., Scambler, P. & et al. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* **328**, 614-616, doi:10.1038/328614a0 (1987).

35 Forrester, K., Almoguera, C., Han, K., Grizzle, W. E. & Perucho, M. Detection of high incidence of K-ras oncogenes during human colon tumorigenesis. *Nature* **327**, 298-303, doi:10.1038/327298a0 (1987).

36 Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E. & Vogelstein, B. The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113, doi:10.1126/science.1145720 (2007).

37 Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson,

J. K., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274, doi:10.1126/science.1133427 (2006).

38 Wang, Z., Shen, D., Parsons, D. W., Bardelli, A., Sager, J., Szabo, S., Ptak, J., Silliman, N., Peters, B. A., van der Heijden, M. S., Parmigiani, G., Yan, H., Wang, T. L., Riggins, G., Powell, S. M., Willson, J. K., Markowitz, S., Kinzler, K. W., Vogelstein, B. & Velculescu, V. E. Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304**, 1164-1166, doi:10.1126/science.1096096 (2004).

39 Bardelli, A., Parsons, D. W., Silliman, N., Ptak, J., Szabo, S., Saha, S., Markowitz, S., Willson, J. K., Parmigiani, G., Kinzler, K. W., Vogelstein, B. & Velculescu, V. E. Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300**, 949, doi:10.1126/science.1082596 (2003).

40 Bass, A. J., Lawrence, M. S., Brace, L. E., Ramos, A. H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A., Jing, R., Parkin, M., Pugh, T., Verhaak, R. G., Stransky, N., Boutin, A. T., Barretina, J., Solit, D. B., Vakiani, E., Shao, W., Mishina, Y., Warmuth, M., Jimenez, J., Chiang, D. Y., Signoretti, S., Kaelin, W. G., Spardy, N., Hahn, W. C., Hoshida, Y., Ogino, S., Depinho, R. A., Chin, L., Garraway, L. A., Fuchs, C. S., Baselga, J., Tabernero, J., Gabriel, S., Lander, E. S., Getz, G. & Meyerson, M. Genomic sequencing of colorectal adenocarcinomas

identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-968, doi:10.1038/ng.936 (2011).

41 Seshagiri, S., Stawiski, E. W., Durinck, S., Modrusan, Z., Storm, E. E., Conboy, C. B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B. S., Guillory, J., Ha, C., Dijkgraaf, G. J., Stinson, J., Gnad, F., Huntley, M. A., Degenhardt, J. D., Haverty, P. M., Bourgon, R., Wang, W., Koeppen, H., Gentleman, R., Starr, T. K., Zhang, Z., Largaespada, D. A., Wu, T. D. & de Sauvage, F. J. Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660-664, doi:10.1038/nature11282 (2012).

42 Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068, doi:10.1038/nature07385 (2008).

43 Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

44 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

45 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

46 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

47 Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

48 Cancer Genome Atlas Research, N., Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B., Kucherlapati, R., Mardis, E. R. & Levine, D. A. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

49 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43-49, doi:10.1038/nature12222 (2013).

50 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

51 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

52 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202-209, doi:10.1038/nature13480 (2014).

53 Cancer Genome Atlas Research, N. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690, doi:10.1016/j.cell.2014.09.050 (2014).

54 Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582, doi:10.1038/nature14129 (2015).

55 Cancer Genome Atlas Research, N., Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., Cooper, L. A., Rheinbay, E., Miller, C. R., Vitucci, M., Morozova, O., Robertson, A. G., Noushmehr, H., Laird, P. W., Cherniack, A. D., Akbani, R., Huse, J. T., Ciriello, G., Poisson, L. M., Barnholtz-Sloan, J. S., Berger, M. S., Brennan, C., Colen, R. R., Colman, H., Flanders, A. E., Giannini, C., Grifford, M., Iavarone, A., Jain, R., Joseph, I., Kim, J., Kasaian, K., Mikkelsen, T., Murray, B. A., O'Neill, B. P., Pachter, L., Parsons, D. W., Sougnez, C., Sulman, E. P., Vandenberg, S. R., Van Meir, E. G., von Deimling, A., Zhang, H., Crain, D., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, T., Sherman, M., Yena, P., Black, A., Bowen, J., Dicostanzo, K., Gastier-Foster, J., Leraas, K. M., Lichtenberg, T. M., Pierson, C. R., Ramirez, N. C., Taylor, C., Weaver, S., Wise, L., Zmuda, E., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M. L., Hutter, C. M., Mills Shaw, K. R., Ozenberger, B. A., Sheth, M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Ayala, B., Baboud, J., Chudamani, S., Jensen, M. A., Liu, J., Pihl, T., Raman, R., Wan, Y., Wu, Y., Ally, A., Auman, J. T., Balasundaram, M.,

Balu, S., Baylin, S. B., Beroukhim, R., Bootwalla, M. S., Bowlby, R., Bristow, C. A., Brooks, D., Butterfield, Y., Carlsen, R., Carter, S., Chin, L., Chu, A., Chuah, E., Cibulskis, K., Clarke, A., Coetzee, S. G., Dhalla, N., Fennell, T., Fisher, S., Gabriel, S., Getz, G., Gibbs, R., Guin, R., Hadjipanayis, A., Hayes, D. N., Hinoue, T., Hoadley, K., Holt, R. A., Hoyle, A. P., Jefferys, S. R., Jones, S., Jones, C. D., Kucherlapati, R., Lai, P. H., Lander, E., Lee, S., Lichtenstein, L., Ma, Y., Maglinte, D. T., Mahadeshwar, H. S., Marra, M. A., Mayo, M., Meng, S., Meyerson, M. L., Mieczkowski, P. A., Moore, R. A., Mose, L. E., Mungall, A. J., Pantazi, A., Parfenov, M., Park, P. J., Parker, J. S., Perou, C. M., Protopopov, A., Ren, X., Roach, J., Sabedot, T. S., Schein, J., Schumacher, S. E., Seidman, J. G., Seth, S., Shen, H., Simons, J. V., Sipahimalani, P., Soloway, M. G., Song, X., Sun, H., Tabak, B., Tam, A., Tan, D., Tang, J., Thiessen, N., Triche, T., Jr., Van Den Berg, D. J., Veluvolu, U., Waring, S., Weisenberger, D. J., Wilkerson, M. D., Wong, T., Wu, J., Xi, L., Xu, A. W., Yang, L., Zack, T. I., Zhang, J., Aksoy, B. A., Arachchi, H., Benz, C., Bernard, B., Carlin, D., Cho, J., DiCara, D., Frazer, S., Fuller, G. N., Gao, J., Gehlenborg, N., Haussler, D., Heiman, D. I., Iype, L., Jacobsen, A., Ju, Z., Katzman, S., Kim, H., Knijnenburg, T., Kreisberg, R. B., Lawrence, M. S., Lee, W., Leinonen, K., Lin, P., Ling, S., Liu, W., Liu, Y., Liu, Y., Lu, Y., Mills, G., Ng, S., Noble, M. S., Paull, E., Rao, A., Reynolds, S., Saksena, G., Sanborn, Z., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Shmulevich, I., Sinha, R., Stuart, J., Sumer, S. O., Sun, Y., Tasman, N.,

Taylor, B. S., Voet, D., Weinhold, N., Weinstein, J. N., Yang, D., Yoshihara, K., Zheng, S., Zhang, W., Zou, L., Abel, T., Sadeghi, S., Cohen, M. L., Eschbacher, J., Hattab, E. M., Raghunathan, A., Schniederjan, M. J., Aziz, D., Barnett, G., Barrett, W., Bigner, D. D., Boice, L., Brewer, C., Calatozzolo, C., Campos, B., Carlotti, C. G., Jr., Chan, T. A., Cuppini, L., Curley, E., Cuzzubbo, S., Devine, K., DiMeco, F., Duell, R., Elder, J. B., Fehrenbach, A., Finocchiaro, G., Friedman, W., Fulop, J., Gardner, J., Hermes, B., Herold-Mende, C., Jungk, C., Kendler, A., Lehman, N. L., Lipp, E., Liu, O., Mandt, R., McGraw, M., McLendon, R., McPherson, C., Neder, L., Nguyen, P., Noss, A., Nunziata, R., Ostrom, Q. T., Palmer, C., Perin, A., Pollo, B., Potapov, A., Potapova, O., Rathmell, W. K., Rotin, D., Scarpace, L., Schilero, C., Senecal, K., Shimmel, K., Shurkhay, V., Sifri, S., Singh, R., Sloan, A. E., Smolenski, K., Staugaitis, S. M., Steele, R., Thorne, L., Tirapelli, D. P., Unterberg, A., Vallurupalli, M., Wang, Y., Warnick, R., Williams, F., Wolinsky, Y., Bell, S., Rosenberg, M., Stewart, C., Huang, F., Grimsby, J. L., Radenbaugh, A. J. & Zhang, J. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England journal of medicine* **372**, 2481-2498, doi:10.1056/NEJMoa1402121 (2015).

56 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).

57 Linehan, W. M., Spellman, P. T., Ricketts, C. J., Creighton, C. J., Fei, S. S., Davis, C., Wheeler, D. A., Murray, B. A., Schmidt, L., Vocke, C. D.,

Peto, M., Al Mamun, A. A., Shinbrot, E., Sethi, A., Brooks, S., Rathmell, W. K., Brooks, A. N., Hoadley, K. A., Robertson, A. G., Brooks, D., Bowlby, R., Sadeghi, S., Shen, H., Weisenberger, D. J., Bootwalla, M., Baylin, S. B., Laird, P. W., Cherniack, A. D., Saksena, G., Haake, S., Li, J., Liang, H., Lu, Y., Mills, G. B., Akbani, R., Leiserson, M. D., Raphael, B. J., Anur, P., Bottaro, D., Albiges, L., Barnabas, N., Choueiri, T. K., Czerniak, B., Godwin, A. K., Hakimi, A. A., Ho, T. H., Hsieh, J., Ittmann, M., Kim, W. Y., Krishnan, B., Merino, M. J., Shaw, K. R., Reuter, V. E., Reznik, E., Shelley, C. S., Shuch, B., Signoretti, S., Srinivasan, R., Tamboli, P., Thomas, G., Tickoo, S., Burnett, K., Crain, D., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J. D., Penny, R. J., Shelton, C., Shelton, W. T., Sherman, M., Thompson, E., Yena, P., Avedon, M. T., Bowen, J., Gastier-Foster, J. M., Gerken, M., Leraas, K. M., Lichtenberg, T. M., Ramirez, N. C., Santos, T., Wise, L., Zmuda, E., Demchok, J. A., Felau, I., Hutter, C. M., Sheth, M., Sofia, H. J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J. C., Zhang, J., Ayala, B., Baboud, J., Chudamani, S., Liu, J., Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Ally, A., Balasundaram, M., Balu, S., Beroukhim, R., Bodenheimer, T., Buhay, C., Butterfield, Y. S., Carlsen, R., Carter, S. L., Chao, H., Chuah, E., Clarke, A., Covington, K. R., Dahdouli, M., Dewal, N., Dhalla, N., Doddapaneni, H. V., Drummond, J. A., Gabriel, S. B., Gibbs, R. A., Guin, R., Hale, W., Hawes, A., Hayes, D. N., Holt, R. A., Hoyle, A. P., Jefferys, S. R., Jones, S. J., Jones, C. D., Kalra, D., Kovar, C., Lewis, L., Li, J., Ma, Y., Marra, M.

A., Mayo, M., Meng, S., Meyerson, M., Mieczkowski, P. A., Moore, R. A., Morton, D., Mose, L. E., Mungall, A. J., Muzny, D., Parker, J. S., Perou, C. M., Roach, J., Schein, J. E., Schumacher, S. E., Shi, Y., Simons, J. V., Sipahimalani, P., Skelly, T., Soloway, M. G., Sougnez, C., Tam, A., Tan, D., Thiessen, N., Veluvolu, U., Wang, M., Wilkerson, M. D., Wong, T., Wu, J., Xi, L., Zhou, J., Bedford, J., Chen, F., Fu, Y., Gerstein, M., Haussler, D., Kasaian, K., Lai, P., Ling, S., Radenbaugh, A., Van Den Berg, D., Weinstein, J. N., Zhu, J., Albert, M., Alexopoulou, I., Andersen, J. J., Auman, J. T., Bartlett, J., Bastacky, S., Bergsten, J., Blute, M. L., Boice, L., Bollag, R. J., Boyd, J., Castle, E., Chen, Y. B., Cheville, J. C., Curley, E., Davies, B., DeVolk, A., Dhir, R., Dike, L., Eckman, J., Engel, J., Harr, J., Hrebinko, R., Huang, M., Huelsenbeck-Dill, L., Iacocca, M., Jacobs, B., Lobis, M., Maranchie, J. K., McMeekin, S., Myers, J., Nelson, J., Parfitt, J., Parwani, A., Petrelli, N., Rabeno, B., Roy, S., Salner, A. L., Slaton, J., Stanton, M., Thompson, R. H., Thorne, L., Tucker, K., Weinberger, P. M., Winemiller, C., Zach, L. A., Zuna, R. & Cancer Genome Atlas Research, N. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *The New England journal of medicine*, doi:10.1056/NEJMoa1505917 (2015).

58 Cancer Genome Atlas Research Network. Electronic address, s. c. m. o. & Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011-1025, doi:10.1016/j.cell.2015.10.025 (2015).

59 Yoshida, K., Sanada, M. & Ogawa, S. Deep sequencing in cancer research. *Japanese journal of clinical oncology* **43**, 110-115, doi:10.1093/jjco/hys206 (2013).

60 Mackenzie, R., Kommoss, S., Winterhoff, B. J., Kipp, B. R., Garcia, J. J., Voss, J., Halling, K., Karnezis, A., Senz, J., Yang, W., Prigge, E. S., Reuschenbach, M., Doeberitz, M. V., Gilks, B. C., Huntsman, D. G., Bakkum-Gamez, J., McAlpine, J. N. & Anglesio, M. S. Targeted deep sequencing of mucinous ovarian tumors reveals multiple overlapping RAS-pathway activating mutations in borderline and cancerous neoplasms. *BMC cancer* **15**, 415, doi:10.1186/s12885-015-1421-8 (2015).

61 Gerlinger, M., Quezada, S. A., Peggs, K. S., Furness, A. J., Fisher, R., Marafioti, T., Shende, V. H., McGranahan, N., Rowan, A. J., Hazell, S., Hamm, D., Robins, H. S., Pickering, L., Gore, M., Nicol, D. L., Larkin, J. & Swanton, C. Ultra-deep T cell receptor sequencing reveals the complexity and intratumour heterogeneity of T cell clones in renal cell carcinomas. *The Journal of pathology* **231**, 424-432, doi:10.1002/path.4284 (2013).

62 Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B. & Loeb, L. A. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513, doi:10.1073/pnas.1208715109 (2012).

63 Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D.,

Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A. & Swanton, C. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

64 de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Gronroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rassl, D. M., Rintoul, R. C., Janes, S. M., Lee, S. M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S. C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P. & Swanton, C. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256, doi:10.1126/science.1253462 (2014).

65 Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C. W., Cao, Y., Gumbs, C., Gold, K. A., Kalhor, N., Little, L., Mahadeshwar, H., Moran, C., Protopopov, A., Sun, H., Tang, J., Wu, X., Ye, Y., William, W. N., Lee, J. J., Heymach, J. V., Hong, W. K., Swisher, S., Wistuba, II & Futreal, P. A. Intratumor heterogeneity in

localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256-259, doi:10.1126/science.1256930 (2014).

66 Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Maner, S., Zetterberg, A., Hicks, J. & Wigler, M. Inferring tumor progression from genomic heterogeneity. *Genome Res* **20**, 68-80, doi:10.1101/gr.099622.109 (2010).

67 Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. *Molecular cell* **58**, 598-609, doi:10.1016/j.molcel.2015.05.005 (2015).

68 Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. & Surani, M. A. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).

69 Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. & McCarroll, S. A. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214, doi:10.1016/j.cell.2015.05.002 (2015).

70 Bose, S., Wan, Z., Carr, A., Rizvi, A. H., Vieira, G., Pe'er, D. & Sims, P. A. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol* **16**, 120, doi:10.1186/s13059-015-0684-3 (2015).

71 Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. & Kirschner, M. W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201, doi:10.1016/j.cell.2015.04.044 (2015).

72 Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J. & Wigler, M. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94, doi:10.1038/nature09807 (2011).

73 Arneson, N., Hughes, S., Houlston, R. & Done, S. Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). *CSH protocols* **2008**, pdb prot4919, doi:10.1101/pdb.prot4919 (2008).

74 Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., Wigler, M., Navin, N. & Hicks, J. Genome-wide copy number analysis of single cells. *Nature protocols* **7**, 1024-1041, doi:10.1038/nprot.2012.039 (2012).

75 Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X. & Wang, J. Single-cell exome sequencing and monoclonal evolution of a JAK2-

negative myeloproliferative neoplasm. *Cell* **148**, 873-885, doi:10.1016/j.cell.2012.02.028 (2012).

76 Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y. & Wang, J. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-895, doi:10.1016/j.cell.2012.02.025 (2012).

77 Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S. F., Egholm, M. & Lasken, R. S. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266, doi:10.1073/pnas.082089499 (2002).

78 Lasken, R. S. Single-cell sequencing in its prime. *Nature biotechnology* **31**, 211-212, doi:10.1038/nbt.2523 (2013).

79 Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626, doi:10.1126/science.1229164 (2012).

80 Chen, K., Meric-Bernstam, F., Zhao, H., Zhang, Q., Ezzeddine, N., Tang, L. Y., Qi, Y., Mao, Y., Chen, T., Chong, Z., Zhou, W., Zheng, X., Johnson, A., Aldape, K. D., Routbort, M. J., Luthra, R., Kopetz, S., Davies, M. A., de Groot, J., Moulder, S., Vinod, R., Farhangfar, C. J., Shaw, K. M.,

Mendelsohn, J., Mills, G. B. & Eterovic, A. K. Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clinical chemistry* **61**, 544-553, doi:10.1373/clinchem.2014.231100 (2015).

81 Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F. & Navin, N. E. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155-160, doi:10.1038/nature13600 (2014).

82 Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

83 Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

84 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).

85 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J. & McCombie, W. R. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522-1527, doi:10.1038/ng.2007.42 (2007).

86 Lasken, R. S. Single-cell genomic sequencing using Multiple Displacement Amplification. *Current opinion in microbiology* **10**, 510-516, doi:10.1016/j.mib.2007.08.005 (2007).

87 de Vega, M., Lazaro, J. M., Mencia, M., Blanco, L. & Salas, M. Improvement of phi29 DNA polymerase amplification performance by fusion of DNA binding motifs. *Proc Natl Acad Sci U S A* **107**, 16506-16511, doi:10.1073/pnas.1011428107 (2010).

88 Leung, M. L., Wang, Y., Kim, C., Gao, R., Jiang, J., Sei, E. & Navin, N. E. Highly multiplexed targeted DNA sequencing from single nuclei. *Nature protocols* **11**, 214-235, doi:10.1038/nprot.2016.005 (2016).

89 Brannon, A. R., Vakiani, E., Sylvester, B. E., Scott, S. N., McDermott, G., Shah, R. H., Kania, K., Viale, A., Oschwald, D. M., Vacic, V., Emde, A. K., Cercek, A., Yaeger, R., Kemeny, N. E., Saltz, L. B., Shia, J., D'Angelica, M. I., Weiser, M. R., Solit, D. B. & Berger, M. F. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol* **15**, 454, doi:10.1186/s13059-014-0454-7 (2014).

90 Tan, I. B., Malik, S., Ramnarayanan, K., McPherson, J. R., Ho, D. L., Suzuki, Y., Ng, S. B., Yan, S., Lim, K. H., Koh, D., Hoe, C. M., Chan, C. Y., Ten, R., Goh, B. K., Chung, A. Y., Tan, J., Chan, C. X., Tay, S. T., Alexander, L., Nagarajan, N., Hillmer, A. M., Tang, C. L., Chua, C., Teh, B. T., Rozen, S. & Tan, P. High-depth sequencing of over 750 genes supports linear progression of primary tumors and metastases in most patients with liver-limited metastatic colorectal cancer. *Genome Biol* **16**, 32, doi:10.1186/s13059-015-0589-1 (2015).

91 Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nature reviews. Genetics* **17**, 175-188, doi:10.1038/nrg.2015.16 (2016).

92 Navin, N. E. Cancer genomics: one cell at a time. *Genome Biol* **15**, 452, doi:10.1186/s13059-014-0452-9 (2014).

93 Miller, M. C., Doyle, G. V. & Terstappen, L. W. Significance of Circulating Tumor Cells Detected by the CellSearch System in Patients with Metastatic Breast Colorectal and Prostate Cancer. *Journal of oncology* **2010**, 617421, doi:10.1155/2010/617421 (2010).

94 Ni, X., Zhuo, M., Su, Z., Duan, J., Gao, Y., Wang, Z., Zong, C., Bai, H., Chapman, A. R., Zhao, J., Xu, L., An, T., Ma, Q., Wang, Y., Wu, M., Sun, Y., Wang, S., Li, Z., Yang, X., Yong, J., Su, X. D., Lu, Y., Bai, F., Xie, X. S. & Wang, J. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc Natl Acad Sci U S A* **110**, 21083-21088, doi:10.1073/pnas.1320659110 (2013).

95 Lohr, J. G., Adalsteinsson, V. A., Cibulskis, K., Choudhury, A. D., Rosenberg, M., Cruz-Gordillo, P., Francis, J. M., Zhang, C. Z., Shalek, A. K., Satija, R., Trombetta, J. J., Lu, D., Tallapragada, N., Tahirova, N., Kim, S., Blumenstiel, B., Sougnez, C., Lowe, A., Wong, B., Auclair, D., Van Allen, E. M., Nakabayashi, M., Lis, R. T., Lee, G. S., Li, T., Chabot, M. S., Ly, A., Taplin, M. E., Clancy, T. E., Loda, M., Regev, A., Meyerson, M., Hahn, W. C., Kantoff, P. W., Golub, T. R., Getz, G., Boehm, J. S. & Love, J. C. Whole-exome sequencing of circulating tumor cells provides a

window into metastatic prostate cancer. *Nature biotechnology* **32**, 479-484, doi:10.1038/nbt.2892 (2014).

96 Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X. S. & Qiao, J. Genome analyses of single human oocytes. *Cell* **155**, 1492-1506, doi:10.1016/j.cell.2013.11.040 (2013).

97 Talmadge, J. E. & Fidler, I. J. AACR centennial series: the biology of cancer metastasis: historical perspective. *Cancer research* **70**, 5649-5669, doi:10.1158/0008-5472.CAN-10-1040 (2010).

98 Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

99 Giuliano, M., Giordano, A., Jackson, S., De Giorgi, U., Mego, M., Cohen, E. N., Gao, H., Anfossi, S., Handy, B. C., Ueno, N. T., Alvarez, R. H., De Placido, S., Valero, V., Hortobagyi, G. N., Reuben, J. M. & Cristofanilli, M. Circulating tumor cells as early predictors of metastatic spread in breast cancer patients with limited metastatic dissemination. *Breast cancer research : BCR* **16**, 440, doi:10.1186/s13058-014-0440-8 (2014).

100 Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M.,

Anders, R. A., Eshleman, J. R., Vogelstein, B. & Diaz, L. A., Jr. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine* **372**, 2509-2520, doi:10.1056/NEJMoa1500596 (2015).

101 Wiener, R. S., Schwartz, L. M., Woloshin, S. & Welch, H. G. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Annals of internal medicine* **155**, 137-144, doi:10.7326/0003-4819-155-3-201108020-00003 (2011).

102 Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P. & Voet, T. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).

VITA

Marco Leung was born in Hong Kong on February 13th, 1988, the son of Timothy Leung and Minnie Yick. He moved to Texas with his family in 2002 and attended John Foster Dulles High School (Class of 2006) in Sugar Land, TX. He received his degree of Bachelor of Science with a major in Molecular Genetic Technology from School of Health Professions at the University of Texas MD Anderson Cancer Center in August 2010. He worked as a research assistant in the Department of Immunology at MD Anderson for one year. In August of 2011, he matriculated at The University of Texas Graduate School of Biomedical Sciences at Houston.