Application of Phylogenetic Analysis in Cancer Evolution

by

Yuantong Ding

Department of Biology
Duke University

Date: __________________________
Approved: ______________________

____________________
Allen Rodrigo, Supervisor

____________________
Paul Magwene, Co-supervisor

____________________
Katharina Koelle

____________________
Charles Nunn

____________________
Edward Patz

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the Graduate School of Duke University

2018
ABSTRACT

Application of Phylogenetic Analysis in Cancer Evolution

by

Yuantong Ding

Department of Biology
Duke University

Date:_______________________

Approved:

___________________________

Allen Rodrigo, Supervisor

___________________________

Paul Magwene, Co-supervisor

___________________________

Katharina Koelle

___________________________

Charles Nunn

___________________________

Edward Patz

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the Graduate School of Duke University

2018
Abstract

Cancer is a major threat to human health and results in nearly one-in-six deaths globally. Despite an extraordinary amount of effort and money spent, eradication or control of advanced disease has not yet been achieved. Understanding cancer from an evolutionary point of view may provide new insight to more effective control and treatment of the disease. Cancer as a disease of dynamic, stochastic somatic genomic evolution was first described by Nowell in 1976, and since then researchers have identified clonal expansions and genetic heterogeneity within many different types of neoplasms. The advancements in sequencing technology, especially single-cell sequencing, has open up new frontiers by bringing the study of genomes to the cellular level. Phylogenetic analysis, which is a powerful tool inferring evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics, has recently been applied to cancer studies and has great promise as a tool for deciphering cancer evolution. However, new challenges have also arisen in experimental design, methodology and interpretation regarding to phylogeny of cancer cells. The overarching theme of this dissertation is to bring phylogenetic analysis to the context of cancer evolution. By using \textit{in silico} simulations, I show the advantages and disadvantages of different sampling designs for phylogenetic analysis. I also develop a new method to infer sub-clone spatial distribution utilizing phased pseudo-haploids from bulk sequencing data. Finally, I
demonstrate the use of phylogenetic analysis in breast cancer with multi-regional bulk sequencing and in lung cancer with single cell sequencing.
Contents

Abstract ........................................................................................................................ iv

List of Tables .................................................................................................................. xi

List of Figures ................................................................................................................ xii

List of Abbreviations .................................................................................................. xiv

List of Terms ................................................................................................................ xv

Acknowledgements ...................................................................................................... xvii

1 Introduction ............................................................................................................... 1

1.1 Cancer evolution .................................................................................................... 1

1.2 Models .................................................................................................................... 3

1.2.1 Clone and sub-clone .......................................................................................... 3

1.2.2 Clonal (truncal), sub-clonal and private mutations ......................................... 5

1.2.3 Monoclonal and polyclonal ............................................................................. 7

1.2.4 Linear and parallel (branched) ......................................................................... 7

1.2.5 Gradual and punctuated (Big Bang) ................................................................. 8

1.2.6 Unified model .................................................................................................... 8

1.3 Experimental technologies and evidences ............................................................ 9

1.3.1 Multi-regional studies ...................................................................................... 10

1.3.2 Single-cell studies ............................................................................................ 11

1.4 Phylogenetic tools .................................................................................................. 16

1.4.1 Sub-clone deconvolution ............................................................................... 16

1.4.2 Phylogenies from single-cell sequencing ......................................................... 16

1.5 Challenges .............................................................................................................. 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>Overview</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>Toward an optimized sequencing experimental design</td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>20</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Bulk sequencing and single-cell sequencing</td>
<td>20</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Comparison of different sequencing designs</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td>Methods</td>
<td>22</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Tumor simulation and lineage tracking</td>
<td>22</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Different sampling methods of tumor cells</td>
<td>26</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Sequence simulation</td>
<td>27</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Latin Hypercube sampling of parameter space</td>
<td>27</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Tree reconstruction and measurements</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>Results</td>
<td>29</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Recovery of the true phylogeny</td>
<td>29</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Measurements profiles of different sampling methods</td>
<td>32</td>
</tr>
<tr>
<td>2.4</td>
<td>Discussion</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>Inference of sub-clone spatial distribution with phylogenies of phased pseudo-haploids</td>
<td>39</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>39</td>
</tr>
<tr>
<td>3.2</td>
<td>Methods</td>
<td>41</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Overview</td>
<td>41</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Simulation of tumor cells</td>
<td>45</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Simulation of sequences</td>
<td>45</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Sampling and mixing</td>
<td>45</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Tree reconstruction and comparison</td>
<td>46</td>
</tr>
</tbody>
</table>
3.2.6 Haploid phasing switch error .......................................................... 47
3.2.7 Compare the sub-clone mixing level in different type of cancer ...... 47
3.3 Results .................................................................................................. 48
3.3.1 Robinson-Foulds distance increases with higher sub-clone mixing level ................................................................................. 48
3.3.2 Adequate mutation rate is required for RF distance to show difference under different sub-clone mixing level ....................... 49
3.3.3 Larger sample size can differentiate sub-clone mixing level on finer scale ...................................................................................... 50
3.3.4 Presence of phasing switch error does not affect the increasing trend of RF distance ................................................................. 51
3.3.5 Empirical data .................................................................................... 52
3.4 Discussion ............................................................................................ 54
4 Method validation and evolutionary analysis of breast cancer, using multi-regional bulk sequencing data ........................................... 58
4.1 Introduction ........................................................................................... 58
4.1.1 Breast cancer ..................................................................................... 58
4.1.2 Ductal carcinoma in situ (DCIS) and Invasive ductal carcinoma (IDC) 59
4.1.3 Metastasis in breast cancer ............................................................... 60
4.1.4 Models of cancer clonal evolution .................................................... 61
4.1.5 Models of invasive cell evolution ..................................................... 64
4.1.6 Evidences for models of invasive cell evolution ............................... 65
4.2 Methods .................................................................................................. 67
4.2.1 Multi-regional bulk whole-exome sequencing in two DCIS cases with invasive component .......................................................... 67
4.2.2 Whole exome sequencing ................................................................ 69
4.2.3 Variant calling...........................................................................................................69
4.2.4 Phylogeny ..................................................................................................................70
4.2.5 Haploid estimation and sub-clone spatial distribution estimation.......................71
4.2.6 Sub-clone estimation ..................................................................................................71
4.2.7 Correlation between genetic distance and physical distance .........................71
4.3 Results ................................................................................................................................72
4.3.1 Phylogeny of SNVs supports multiple origins of invasive cells .......................72
4.3.2 Invasive cells in primary tumor and lymph nodes may follow different origin models..................................................................................................................74
4.3.3 Timing of invasive cells and mutations .................................................................75
4.3.4 Sub-clone and spatial distribution...........................................................................75
4.3.5 Significant correlation between genetic distance and physical distance.............77
4.3.6 Mutations in potential driver genes .......................................................................77
4.4 Discussion .....................................................................................................................78
5 Validation and interpretation of single-cell sequencing of lung cancer.................81
5.1 Introduction ....................................................................................................................81
5.1.1 Lung cancer ..............................................................................................................81
5.1.2 Non-small cell lung cancer (NSCLC) .....................................................................82
5.1.3 Prognostic markers ..................................................................................................83
5.1.4 Intra-tumor heterogeneity in lung cancer ...............................................................84
5.2 Methods .........................................................................................................................86
5.2.1 Data set ....................................................................................................................86
5.2.2 Mutation calling ........................................................................................................86
List of Tables

Table 1: Summary of single-cell studies. ................................................................. 14
Table 2: Parameters range in Latin hypercube sampling. ........................................ 26
Table 3: Tree balance statistics comparison of true lineage and reconstructed ones. .... 34
Table 4: Mantel’s test .................................................................................................. 35
Table 5: Parameter settings in simulations ................................................................ 46
Table 6: Tree statistics ................................................................................................. 89
List of Figures

Figure 1: Genetic heterogeneity in different cancers......................................................... 2
Figure 2: Difference between tumor clones and unicellular species ................................. 4
Figure 3: Terminologies in cancer evolution and phylogeny............................................. 6
Figure 4: Punctuated equilibrium framework. ..................................................................... 9
Figure 5: Illustration of incomplete sampling of primary tumor sub-clones results in inferring wrong mode of metastasis. ................................................................. 18
Figure 6: Flowchart and sampling scheme of simulation in Chapter 2. .............................. 25
Figure 7: Distribution of scaled symmetric difference between true and reconstructed lineages................................................................................................................................. 30
Figure 8: Examples of true and reconstructed lineages with different sampling methods. ................................................................................................................................................. 31
Figure 9: Average cell type composition of single-cell samples obtained with different sampling methods. ................................................................................................................. 35
Figure 10: Illustration of sub-clone spatial distribution and phylogeny............................. 43
Figure 11: RF distance under different population size and sample size. .......................... 49
Figure 12: Effect of error rate on RF distance under different mutation rate. .................... 52
Figure 13: Phylogenies based on each set of estimated haploids in colon, liver and breast tumor .............................................................................................................................. 53
Figure 14: Illustration of DCIS and invasive breast tumor. .................................................. 60
Figure 15 Linear and branched evolution pattern................................................................ 62
Figure 16: Models of cancer clonal evolution. ..................................................................... 63
Figure 17: Evolutionary models of invasions in DCIS (140). .............................................. 65
Figure 18: Tumor dissection and sampling locations. ......................................................... 68
Figure 19: Phylogenetic trees. ............................................................................................. 73
Figure 20: Deconvolution of sub-clone. ................................................................. 74
Figure 21: Trees built with estimated haploid sets. ........................................... 76
Figure 22: Linear regression of physical distance and genetic distance............. 77
Figure 23: Heat map of somatic mutations on potential driver genes............. 78
Figure 24: Types of lung cancer.......................................................................... 83
Figure 25: Maximum-Likelihood tree of 12 single cells in one patient and 1 cell-line control. .................................................................................................................. 89
Figure 26: Heatmap of potential driver mutations and mutation tree. .............. 91
Figure 27: Estimation of effective population size........................................... 92
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADO</td>
<td>Allele dropout</td>
</tr>
<tr>
<td>CTC</td>
<td>Circulating tumor cell</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal carcinoma in situ</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen receptor</td>
</tr>
<tr>
<td>HER2</td>
<td>Human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>IDC</td>
<td>Invasive ductal carcinoma</td>
</tr>
<tr>
<td>IE ratio</td>
<td>Ratio between internal and external branch length</td>
</tr>
<tr>
<td>ITH</td>
<td>Intra-tumor heterogeneity</td>
</tr>
<tr>
<td>LHS</td>
<td>Latin Hypercube Sampling</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>NJ</td>
<td>Neighbor-Joining</td>
</tr>
<tr>
<td>NSCLC</td>
<td>Non-small cell lung cancer</td>
</tr>
<tr>
<td>PD</td>
<td>Phylogenetic diversity</td>
</tr>
<tr>
<td>PR</td>
<td>Progesterone receptor</td>
</tr>
<tr>
<td>RF</td>
<td>Robinson-Foulds metric</td>
</tr>
<tr>
<td>SCS</td>
<td>Single-cell sequencing</td>
</tr>
<tr>
<td>SNV</td>
<td>Single-nucleotide variation</td>
</tr>
</tbody>
</table>
List of Terms

Definitions of cancer terminology according to American Cancer Society(1) and American Society of Clinical Oncology(2):

**Adenocarcinoma**: cancer that starts in glandular tissue, such as in the ducts or lobules of the breast or in the gland cells of the prostate.

**Benign**: refers to a tumor that is not cancerous (not malignant). The tumor does not usually invade nearby tissue or spread to other parts of the body.

**Cancer**: a group of diseases which cause cells in the body to change and grow out of control. Most types of cancer cells form a lump or mass called a tumor. (Not all tumors are cancer. A tumor that is not cancer is called benign, while a cancerous tumor is called malignant.)

**Carcinoma**: a cancer that begins in the lining layer (epithelial cells) of organs.

**In situ**: In place. Refers to cancer that localized and confined to one area, not spread to nearby tissue, also called non-invasive cancer.

**Invasive cancer**: cancer that has spread beyond the layer of cells where it first began and has grown into nearby tissues.

**Lymph nodes**: tiny, bean-shaped organs that help fight infection. Part of the lymphatic system. Through the lymphatic system, cancer can spread to other parts of the body.
**Malignant**: cancerous, refers to a tumor that is cancerous. It may invade nearby healthy tissue or spread to other parts of the body.

**Metastasis**: the spread of cancer from the place where the cancer began to another part of the body, often by way of the lymph system or bloodstream. Regional or local metastasis is cancer that has spread to the lymph nodes, tissues, or organs close to where the cancer started (the primary site). Distant metastasis is cancer that has spread to organs or tissues that are farther away.

**Neoplasm**: an abnormal growth (tumor) that starts from a single altered cell. A neoplasm may be benign (noncancerous) or malignant (cancerous).

**Squamous cell carcinoma**: cancer that begins in the flat, non-glandular cells of the body, for example, the skin or the lining of the body’s organs.

**Triple-negative breast cancer**: breast cancer that does not have three hormone receptors: estrogen receptors, progesterone receptors and human epidermal growth factor receptor 2 (HER2). These hormones would fuel cell grows. Triple-negative patients will not benefit from targeted therapies which block hormone receptors.

**Tumor**: an abnormal lump or mass of tissue. Tumors can be benign (noncancerous) or malignant (cancerous, meaning it can spread to other parts of the body). Also called a nodule or mass.
Acknowledgements

I would like to thank my wonderful advisor Allen Rodrigo and co-advisor Paul Magwene for encouraging and guiding me through my graduate school and beyond. They have shown me, by their example, what a good scientist and person should be.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects: my committee members Katia Koelle, Charles Nunn, Edward Patz, former committee member Jianguo Liu and collaborator Marc Ryser. They have all inspired me for new ideas and provided me valuable professional guidance. And thanks Dr. Shelley Hwang’s group and Dr. Patz’s group for providing the empirical data in this dissertation.

I would like to thank the Biology Department support staff in the Duke Biology Department, Duke Compute Cluster and the Research School of Biology at the Australian National University. And I would also like to thank Triangle Center for Evolutionary Medicine (TriCEM) for funding.

Finally, thank you also to the many people who have provided me with support and friendship over the years: lab members Steven Wu, Qinglong Zeng, people I met in Australian National University Thomas Wang, Louis Ranjard, Xia Hua, and fellow students in Duke Biology. They all made this journey full of joy and surprises.
1 Introduction

1.1 Cancer evolution

Cancer is the second leading cause of death globally after heart disease, and was responsible for 8.8 million deaths in 2015 (3). Globally, nearly 1 in 6 deaths is due to cancer (3). Despite an extraordinary amount of effort and money spent and a much greater understanding of cancer biology and genetics (4), eradication or control of advanced disease has not been achieved. Understanding the cellular complexity of the disease, its dynamic and evolutionary characteristics may provide new insight to more effective control and treatment of the disease (5). From an evolutionary perspective, cancers are thought to arise when somatic genomic instability generates cellular variants that differ in their abilities to secure resources and to proliferate. Cells that compete successfully, survive and persist, are the evolutionary precursors of cancer (5).

The notion of cancer as an evolutionary process was first described by Nowell in 1976 (6). He described the model for evolution of tumor cell populations as a unicellular-originated and stepwise clonal growth pattern, meaning tumor cells initiate from a single normal cell. These cells accumulate mutations sequentially, some of which are selectively advantageous. Since mutations accumulate during replication, and therefore, provide the raw material upon which selection can act, genetic heterogeneity is expected in a tumor. However, it is only with the advent of next-generation sequencing studies that the full extent of genomic intra-tumor heterogeneity (ITH) is
becoming apparent. Currently, genetic heterogeneity has been confirmed in many different types of neoplasms, including esophageal adenocarcinoma(7), chronic lymphocytic leukemia(8, 9), acute myeloid leukemia(10, 11), renal-cell carcinoma(12), colorectal carcinogenesis(13, 14), lung adenocarcinomas (15, 16), glioblastoma(17), breast cancer(18-20), bladder cancer(21) and hepatocellular carcinoma(22). A summary of ITH proportion across different types of tumor is shown in Figure 1(23).

Figure 1: Genetic heterogeneity in different cancers. For each tumor type, each point represents one tumor, with the proportion of heterogeneous mutations (ITH proportion), as well as the absolute numbers of
heterogeneous and homogeneous non-silent mutations, shown. Black circles represent treatment naive tumors, with red triangles indicating tumors that have received treatment. Notably, these data are restricted to non-silent mutations and does not include copy-number alterations. (Figure cited from Navin et.al 2017(24))

Understanding cancer as an evolutionary process has profound implications for prevention (25) and treatment (26) alike. And ITH is considered as the key factor contributing to the lethal outcome of cancer, therapeutic failure, and drug resistance (27). Cancerous cells with mutations that confer resistance to anti-cancer therapies are likely to promote tumor persistence and therapeutic resistance (28). For example, Ding and colleagues showed that acute myeloid leukemia cells routinely acquire a small number of additional mutations at relapse which contribute to chemotherapy resistance (29). Conversely, Kostadinov and colleagues showed that non-steroidal anti-inflammatory drugs (NSAIDs), which can reduce cancer risk, might function by reducing somatic genomic abnormality acquisition rate (30).

1.2 Models

To understand the evolution of a tumor, we need to develop a model of how cancer cells proliferate. But before we discuss different models, we need to define a few key terms.

1.2.1 Clone and sub-clone

Clone and sub-clone may have different definitions based on context. In general, there are three main definitions. The first one defines clone in a phylogenetic context, and are used synonymously with “clade” and “sub-clade”. This concept was first proposed in Nowell’s 1976 paper(6), which is a parallel to Darwinian natural selection,
with cancer clones as the equivalent of asexually reproducing, unicellular quasi-species (5).

The second is defined in a phonetic context, and are based on genotypic and phenotypic similarity (31). For example, all tumor cells may be considered as one clone, with possible multiple sub-clones.

The third common definition of clone is based on whether they share the same sets of driver mutations. In this definition, clones may designate groups of cells that correspond to ancestral genotypes (defined by their complement of driver genes), and not just to terminal nodes on the genealogy. This contrasts with the idea of clone/sub-clones as terminal nodes on a phylogeny. The difference is illustrated in Figure 2.

As pointed out by Quake and colleagues (32), the use of clone and sub-clone structure is ambiguous and inconsistent among current studies. Therefore, it’s important to examine these definitions when interpreting and comparing results from
different studies. In this dissertation, to keep congruency with currently literature, the meaning of clone/sub-clone may differ depending on context and will be explained accordingly.

1.2.2 Clonal (truncal), sub-clonal and private mutations

Clonal, sub-clonal and private are often used to describe the frequency of certain mutations or mutated gene in samples (Figure 3). If one mutation or one gene is mutated in all samples, it’s described as clonal and also may be referred to as truncal. If it’s shared by some but not all samples, it’s sub-clonal. And if it’s unique to one specific sample, it’s private. The equivalent terms used in phylogeny for truncal, sub-clonal and private mutations are symplesiomorphies of tumor cells (mutations that occur only in tumor clade), synapomorphies (occur only in some tumor clades), and autapomorphies (occur only in one cell), respectively. They are illustrated in Figure 3. These terms may help us to identify key genes in tumor progression and potentially targeted treatment design.
Figure 3: Terminologies in cancer evolution and phylogeny.
1.2.3 Monoclonal and polyclonal

These terms are often used to describe the origin of certain cells of interest (primary tumor cells, metastasis or cells with mutations on specific gene) (Figure 3). Monoclonal and polyclonal refers to whether the cells of interest have one or multiple origins. The equivalent phylogenetic terms are monophyletic and polyphyletic. The cells of interest could be all tumor cells or metastasis cells. Both clustering and phylogeny have been used to investigate the monoclonal or polyclonal pattern.

1.2.4 Linear and parallel (branched)

These terms are similar to monoclonal and polyclonal but with a focus on general evolutionary pattern of all cells rather than cells of interest (Figure 3). The term “linear” means that mutations are acquired sequentially and fix in the population quickly and persist in subsequent lineages of tumor cells. In contrast, “parallel” or “branched” means that cells (and lineages) with different driver mutations can co-exist (33). Linear models generate unbalanced/ladder-like phylogenies while parallel models generate more balanced phylogenies. A linear model is likely to occur when there are sequential selective sweeps(34). However, evidence of linear branching model based on histological staining, methylation analysis, or PCR genotyping of specific gene are likely underestimated the heterogeneity level in tumor(24). In general, the evidence in support of linear models in cancer is limited(24).
1.2.5 Gradual and punctuated (Big Bang)

These terms describe the timing of mutations arising (Figure 3). According to Navin and colleagues’ definition, “gradualism” means that mutations that induce measurable phenotypic changes are acquired sequentially(33). “Punctuation” refers to the phenomenon where infrequent periods of genomic instability generate so-called catastrophic mutations that cause dramatic phenotypic change are interspersed with relatively long period of genomic/phenotypic stasis (33). This is also the key character of Big Bang model (35). Equivalent concepts in evolutionary biology are phyletic gradualism and punctuated equilibrium, the concepts that morphological difference between species accumulates gradually over time or in a short burst, respectively. Punctuated mutational mechanisms, especially copy number variation, are complex and very common in cancer, but have an as yet undefined role in transformation(36).

1.2.6 Unified model

A hybrid model based on punctuated equilibrium(37) offers a framework that includes the concepts discussed above(33). It defines the cellular species based on similarities in phenotypes that are relevant to the hallmarks of cancer(33). In this model, tumor populations can undergo a series of different phases: punctuation, gradualism and stasis (Figure 4). The punctuation phase is a period of high genomic instability that causes dramatic phenotypic change and a rapid evolutionary tempo. The gradualism phase is a period of sequential accumulation of mutations in the pre-malignancy phase. Stasis phase is a period of neutral accumulation of passenger mutations during tumor
growth. This model includes the general pattern of lineages, mutation timing and pattern. However, it does not explicitly state the monoclonal or polyclonal origin of primary or metastatic tumor cells.

![Punctuated equilibrium framework](image)

**Figure 4:** Punctuated equilibrium framework. (cited form adopted from Graham and Weight (33))

### 1.3 Experimental technologies and evidences

Several experimental strategies have been used to investigate within-tumor evolutionary processes, but one common feature is to have multiple samples from an individual tumor. The sampling resolutions could be multi-regional bulk samples, single-cell samples or combinations of single-cell and bulk samples. The technologies used includes low-coverage sequencing, high-coverage sequencing, deep-sequencing or a
combination of different methods. The sampling site could be the primary tumor, metastases, recurrent sites or circulating tumor cells. Targeted regions could be whole genomes, whole exomes, targeted gene panels or specific genes. The data type obtained could be copy number variations (CNVs), single nucleotide variants (SNVs) or both. Here we provide a review of the current major multi-regional bulk and single-cell DNA sequencing studies.

1.3.1 Multi-regional studies

In this section, I review some recent results obtained from multi-regional sampling studies from a range of cancers.

In breast cancer, a multi-regional sequencing study of 303 samples from 50 patients found parallel evolution for driver genes, whereby different mutations affecting the same gene occur in parallel on different branches (38). They also found evidence for the early origins of metastatic sub-clones and for the multi-clonal invasion model, whereby different metastases originate from different clones of the primary tumor. Similarly, in colon cancer, metastasis can arise from either early or late lineages in primary tumor (39).

In prostate cancer, one longitudinal study from four patients with samples obtained from primary and metastatic tumors revealed multiple waves of metastatic seeding from the primary tumor (40).

Zhao et al. used bulk exome sequencing for 32 primary and 139 matched metastatic tumor tissues in the same patient from 40 subjects in 13 types of cancer.
They found metastases can originate from multiple primary tumor lineages and can arise early in tumor development\(^{(41)}\).

1.3.2 Single-cell studies

With bulk sequencing, the sensitivity of detection is limited to variants that are present in about 20% of cells of a bulk sample composed of thousands of cells\(^{(32)}\). Thus single-cell genomics can open up new frontiers by bringing the study of genomes to the level of the individual cell \(^{(32)}\). A summary of current single-cell sequencing studies is listed in Table 2.

In breast cancer, monoclonal\(^{(42-44)}\) pattern for primary tumor has been identified with single-cell copy number variation studies\(^{(42-44)}\) and single-cell targeted sequencing\(^{(42)}\). A linear progression pattern was found in one primary tumor study \(^{(42)}\). Invasion in breast cancer has been found to have both monoclonal and polyclonal origins with CNV data \(^{(42, 44)}\). For the timing or mutations, chromosome rearrangements are thought to occur early, in punctuated bursts of evolution, while point mutations appear to accumulate gradually \(^{(43)}\). Mutation rate in a triple-negative tumor (tumor that have no or few of the three hormone receptors \(ER, PR\) and \(HER2\), which is important for targeted drug treatment) was estimated to have a 13-fold increase, whereas a ER+ tumor cells did not have an appreciable increase in mutation rate\(^{(43)}\).

In colon cancer, both monoclonal\(^{(45)}\) and polyclonal\(^{(45, 46)}\) patterns for primary tumor were found with single-cell SNV\(^{(45, 46)}\) and CNV\(^{(45)}\) data. Both patterns
for metastases were also found (45). Additionally, metastases were thought to arise late from the primary tumor (45).

In leukemia, tumor cells originate in bone marrow. Evidence from single-cell SNV data showed that tumors could be either monoclonal or polyclonal in acute myeloid leukemia at the time of diagnosis of myelodysplastic syndrome (47). There is also evidence for monoclonal evolution in myeloproliferative neoplasms (48). The branching architecture of sub-clones has been found in multiple single-cell SNV studies in acute myeloid leukemia and multiple myeloma (47, 49, 50). Evidence exists for punctuated structure variation in acute lymphoblastic leukemia (51).

In summary, current multi-regional studies have tended to focus on the relationship between metastases and the primary tumor while single-cell studies capture more details of primary tumor evolution. Monoclonal (monophyletic) pattern for primary tumor was identified in breast cancer and leukemia by single-cell studies. Both monoclonal (monophyletic) and polyclonal (polyphyletic) origins for primary tumor has been found in colon cancer with single-cell studies. A branching pattern is commonly observed in both single-cell and multi-regional studies across different type of cancers. Both monoclonal (monophyletic) and polyclonal (polyphyletic) origins for invasive cells are supported by single-cell studies in breast cancer and colon cancer, and multi-regional sequencing in breast cancer. Invasive cells were observed to arise from either early, late or at different time point from primary tumor. Although phylogenetic approaches were often used in these studies to address tumor evolution, these studies
employ different methods for building trees including clustering and different methods for phylogenetic inference. Whether they are directly comparable is unclear.

Furthermore, single-cell sequencing is still in its infancy, and has only been tested in a limited number of tumors and cases as proof-of-principle studies. Further methodological development and data accumulation are needed to understand the application and value of these methods, as they relate to our understanding of tumor evolution.
### Table 1: Summary of single-cell studies.

<table>
<thead>
<tr>
<th>studies</th>
<th>Disease</th>
<th># sampled single cells</th>
<th># patients</th>
<th>Method</th>
<th>Data</th>
<th>Support</th>
<th>Phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navin et al. 2011</td>
<td>Breast cancer</td>
<td>200</td>
<td>2</td>
<td>Whole genome</td>
<td>CNV</td>
<td>Monoclonal for both primary and metastasis</td>
<td>NJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sequential clonal expansion</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Punctuated clonal expansions, late seeding of metastasis</td>
<td></td>
</tr>
<tr>
<td>Xu et al. 2012</td>
<td>Kidney cancer</td>
<td>25</td>
<td>1</td>
<td>Exom seq</td>
<td>SNV</td>
<td>Branching, no obvious cell subpopulations</td>
<td>PCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Punctual for primary tumor</td>
<td></td>
</tr>
<tr>
<td>Hou et al. 2012</td>
<td>hematopoietic tumors</td>
<td>58</td>
<td>1</td>
<td>Exome seq</td>
<td>SNV</td>
<td>Monoclonal for primary tumor</td>
<td>PCA, clustering</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hematopoietic tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Li et al. 2012</td>
<td>Bladder cancer</td>
<td>66</td>
<td>1</td>
<td>Exome</td>
<td>SNV</td>
<td>Monoclonal for primary tumor</td>
<td>PCA, clustering</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Branching</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>GJ Morgan et al. 2014</td>
<td>Multiple myeloma</td>
<td>73-243</td>
<td>6</td>
<td>qPCR</td>
<td>genotyping</td>
<td>Branching, parallel for driver genes</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TA Graubert et al. 2014</td>
<td>Acute myeloma</td>
<td>36</td>
<td>3</td>
<td>qPCR</td>
<td>SNV</td>
<td>Branching</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yu et al. 2014</td>
<td>Colon cancer</td>
<td>63</td>
<td>1</td>
<td>Exome seq</td>
<td>SNV</td>
<td>Bi-clonal origin of primary tumor</td>
<td>UPGMA</td>
</tr>
<tr>
<td>Study</td>
<td>Tumor Type</td>
<td>Sample Size</td>
<td>Methodology</td>
<td>Genes Assessed</td>
<td>Genetic Changes</td>
<td>Mutational Patterns</td>
<td>Clonal Evolution</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>--------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Leung et al. 2018</td>
<td>Colon cancer</td>
<td>74(CNV), 372(1000 gene panel)</td>
<td>2 mixed</td>
<td>CNV, SNV</td>
<td>both monoclonal and polyclonal pattern (primary and metastasis)</td>
<td>-</td>
<td>Late dissemination of metastasis</td>
</tr>
<tr>
<td>Wang et al. 2014</td>
<td>Breast cancer</td>
<td>100(CNV), 20(WGS), 75(WES)</td>
<td>2 mixed</td>
<td>CNV, SNV</td>
<td>monoclonal</td>
<td>-</td>
<td>CNV occur early, punctuated burst, SNV gradual</td>
</tr>
<tr>
<td>Hughes et al. 2014</td>
<td>Acute myeloid leukemia</td>
<td>36 WGS</td>
<td>3 SNV</td>
<td>-</td>
<td>branching</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gawad et al. 2014</td>
<td>Acute lymphoblastic leukemia</td>
<td>1479 Targeted gene panel</td>
<td>6 SNV</td>
<td>-</td>
<td>-</td>
<td>Punctuated structure variation</td>
<td>-</td>
</tr>
<tr>
<td>Hou et al. 2017</td>
<td>Hepatocellular carcinoma</td>
<td>25 Triple seq</td>
<td>1 CNV, DNA methylome, transcriptome</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Clustering of CNV</td>
</tr>
</tbody>
</table>
1.4 Phylogenetic tools

1.4.1 Sub-clone deconvolution

For DNA sequencing data obtained from bulk samples, inferring sub-clonal structure alone has been addressed by first calling SNVs and then grouping the SNVs into clones according to their estimated frequencies using mixture models. PyClone [26], THetA [27], SciClone(53) and subcloneSeeker(54) all follow this strategy and also correct SNV frequencies for copy number variations.

There are also methods to directly infer both sub-clones and their phylogenetic relationships from bulk sequencing data. A unified approach to combine clustering and tree inference with clones at inner nodes is PhyloSub(55), which is based on non-parametric Bayesian mixture modeling using Dirichlet process. BitPhylogeny(56) is a non-parametric Bayesian approach that jointly estimates the number and composition of clones in the sample as well as the most likely tree connecting them. PhyloWGS(57) reconstructs complete genotypes of these subpopulations based on variant allele frequencies (VAFs) of point mutations and population frequencies of structural variations. The recently published suite of tools, Treeomics(58) uses a Bayesian approach to estimate mutational patterns, and then find the most reliable and evolutionarily compatible mutation pattern for all variants.

1.4.2 Phylogenies from single-cell sequencing

Single-cell-based studies for deriving tumor phylogenies rely on single-cell somatic single-nucleotide variation (SNV) profiles. Distance-based methods like UPGMA
and neighbor joining have been used by Yu et al.(46) and Xu et al(59). Model-based methods like Maximum Likelihood has also been used(47).

There are several methods for inferring tumor phylogenies that take single-cell sequencing specific error model into consideration. The prominent single-cell WGA errors include: allelic dropout (ADO) errors, false positive errors (FPs), non-uniform-coverage distribution, and low-coverage regions(60). ADO is responsible for falsely representing the heterozygous genotypes as homozygous and contributes a considerable number of false negatives (i.e., true SNVs that are not identified) in point mutations.

SCITE(61), OncoNEM(62) and SiFit(63) are three new tools that were specifically designed for inferring tumor evolution from single-cell sequencing data. SCITE is a Markov chain Monte Carlo (MCMC) algorithm that allows one to infer a maximum likelihood (ML) tree from the imperfect genotype matrix. It infers the evolutionary history as a mutation tree(64), which shows the chronological order of the mutations that occur during tumor development. OncoNEM is a likelihood-based method that employs a heuristic search algorithm to find the ML clonal tree, a condensed tree that represents the evolutionary relationship between the subpopulations in the data. Both methods use infinite sites assumption. The third tool SiFit is also likelihood-based but uses finite sites model.
1.5 Challenges

Although phylogenetic methods have shown great potential to address important questions about tumor progression, there are also confounding factors such as healthy contamination, un-sampled/extinct clones, selection, or rate variation among lineages when apply to tumor data.

Sampling strategies is one of the factors that may affect phylogeny inference and interpretation. For example, incomplete sampling of the primary tumor may lead to incorrect inference of the metastatic chronology(65). As illustrated in Figure 5, the true lineages show three parallel metastases from the primary tumor. However, due to incomplete sampling, the inferred tree generates a monophyletic clade of metastases.

![Figure 5: Illustration of incomplete sampling of primary tumor sub-clones results in inferring wrong mode of metastasis. Red dots represent metastatic cells. The metastatic cells have polyclonal origin (polyphyletic) in true lineages. However, due to incomplete sampling, the inferred phylogeny shows a monoclonal (monophyletic) pattern.](image)

Tumor data specific phylogenetic tools has been developed in recent years but there are rooms for further improvement. Ideally, these tools should be able to compare multiple normal and matched tumor (primary or metastatic) samples, infer sub-clones from bulk sequencing data, building phylogenies based on genotype data,
annotate potential driver genes along the inferred trees, incorporate spatial information and provide confidence measurements. However, to date, no single framework achieves all these aims. There is large potential for development of more diverse tumor specific phylogenetic tools

1.6 Overview

The overarching theme of this dissertation is to explore the application of phylogenetic analyses to the study of cancer evolution and developing novel phylogenetic methods addressing questions in cancer. In particular, I am interested how our sampling designs affect our inferences. I address this in Chapter 2, when I explore the optimal sampling strategies for within-tumor sequencing studies with simulations. If there are constraints in sampling, and only data from bulk samples can be obtained, there may still be interesting and novel methods to extract evolutionary information. This is the basis of Chapter 3, in which I present a new phylogeny-based method to infer the sub-clonal spatial distribution within tumors given bulk sequencing data. In Chapter 4 and 5, I focus on analyzing two empirical data sets of multi-regional bulk sequencing in breast cancer and single-cell sequencing in lung cancer.
2 Toward an optimized sequencing experimental design

2.1 Introduction

Multi-regional bulk sequencing and single-cell sequencing have been used to build phylogenetic trees and infer the evolutionary process in tumor. However, whether phylogenies based on these different types of data would differ is unclear. In this chapter, I use simulations to explore the effect of sampling strategies on phylogeny.

2.1.1 Bulk sequencing and single-cell sequencing

As noted in Chapter 1, the notion of cancer as an evolutionary process was first described by Nowell in 1976 (6). Although the idea of cancer evolution is not new, characterization of the processes at the genomic level has only become feasible with the advent of recent technologies, in particular next generation sequencing (NGS). However, methodological challenges remain with respect to the analysis of such data.

Importantly, because traditional bulk sequencing methods are not designed to resolve genetic heterogeneity in mixed populations of cells, standard analyses are subject to a substantial loss of phylogenetic information.

The recent development of single-cell sequencing technology provides a formidable solution to this problem and enables the reconstruction of phylogenies based on samples of single cells within tumors (49). It has the potential to revolutionize the way we view cancer and the way we approach it from both the experimental and clinical perspectives (66). To date, single-cell DNA sequencing studies of tumors has been performed in multiple myeloma (50), breast cancer (42, 43), bladder cancer (52),
myeloproliferative neoplasm (48), leukemia (49, 67), lung cancer (68) and clear cell renal cell carcinoma (59). These studies use single cells obtained from solid tumor tissue (42, 43, 52, 59) or bone marrow aspirates and peripheral blood (48-50, 67, 68), with subsequent isolation using microfluidic (68), flow cell sorting based on nuclei content (42, 43, 49, 50, 67) or cascade dilution (48, 52, 59).

2.1.2 Comparison of different sequencing designs

A key issue in the use of single-cell sequencing (SCS) is how to best sample cells from a solid tumor, particularly given the widespread cellular genomic heterogeneity is spatially distributed across the tumor (12, 15, 17, 67, 69, 70). Simulation studies provide a means to study the question of optimal sampling schemes. By combining a computational model of solid tumor growth that keeps track of genealogy of cells with phylogenetic inference methods, the performance of alternative sampling schemes can be directly evaluated. To our knowledge, and with the exception of an agent-based model of colon cancer that includes lineage tracking of cells along the one-dimensional linear array of colon crypts (71), little work has been done to address this issue. However, the crypt structure is unique to colon cancer, which makes the previously published model inapplicable to other solid tumors.

Our objective was to utilize a two-dimensional discrete model to simulate the tumor development on a single cell level and to explore through simulation how different sampling strategies affect phylogenetic estimation. We focused on alternative single-cell sampling schemes with respect to sampling location within a solid tumor, and
we compared the performance of single-cell against conventional bulk sampling approaches.

2.2 Methods

2.2.1 Tumor simulation and lineage tracking

To simulate the growth of a solid tumor, we developed a spatial model of cell division, cell death and mutation in two spatial dimensions. More precisely, cells are placed on a regular two-dimensional grid and each node is either empty or occupied by at most one cell. When one cell dies, its site becomes unoccupied, and upon cell division, one of the daughter cells competes for space with one randomly chosen cell among the 8 nearest neighbors. There are four distinct cellular phenotypes: normal cells and three different tumor cells types of increasing malignancy, corresponding to successive accumulation of driver mutations. Each cell type has its proliferation rate, fitness, transition probability (probability of acquiring a driver mutation which results in transition to a more malignant cell type), death rate and rate of neutral passenger mutations (nucleotide substitution rate). The parameters are defined as follows (Table 3):

**Proliferation rate** – the probability of a chosen cell to proliferate at each time step

**Death rate** – the probability of cell death at a certain time step
**Transition probability** – the probability of cells of a certain type to transit to the next type (can be considered as the probability of driver mutation occurrence)

**Fitness** – the competitive advantage between different types of cells, probability of cell i replacing cell j is given by \( \frac{f_i}{f_i + f_j} \).

**Passenger mutation accumulation rate (substitution rate)** – background nucleotide substitution rate

Our simulation begins with a single tumor cell in a square 300×300 grid filled with normal cells. All cancer cells are descended from this initial tumor cell. Since we are concerned with the genealogy of all cells in a sample, we also need a mechanism to obtain the genealogy of normal cells that may appear in the sample. To do this, we have a pre-simulation step, where we create a field of 1000×1000 normal cells, starting from a single cell. By pre-simulating, and keeping track of, a forward-time genealogy of all normal cells in this field, we are able to select a sub-field of 300 x 300 cells of known genealogy to seed the simulation of tumor growth.

During primary tumor simulation, we used Gillespie algorithm(72) to model the birth-death process of cells as follows:

1. Initialize the system
2. Generate waiting times for the next birth/death event for each type of cells
3. Find the minimum time and corresponding event then update the system
4. Return to Step 2 until stopping criteria has been satisfied
At each birth event a cell is randomly picked to replace its neighbor with its offspring. The flowchart of cell behavior is shown in Figure 6A. During the primary tumor simulation, the lineages of cells are also tracked. This allows a complete genealogy of any subsample of cells to be retrieved at any time-step of the simulation. The tumor simulation stops when the first tumor cell hits the boundary of the simulation domain. At this point, 50 single cells or groups of cells are chosen in accordance to the sampling method (Figure 6B).
Figure 6: Flowchart and sampling scheme of simulation in Chapter 2. a. Flowchart of cellular dynamics. Each cell can proliferate or die as determined by the Gillespie algorithm (72). If proliferation is chosen, the parental cell will divide to two offspring cells. If proliferation is chosen, the parental cell will divide to two offspring cells. One will replace the parental cell and the other offspring will compete with a random neighbor. It can replace the neighbor with a probability given by its fitness. Both offspring cells can mutate to the next type. b. Sampling location and modality. There are 4 different sampling locations. Random sampling is scattered sampling across the entire (two-dimensional cross-section of the) tumor. Core sampling samples only from the core area, which has the size of half of the total tumor area. Boundary sampling samples from areas other than the core. Section sampling samples evenly on a line starting from the tumor center to the boundary. For each of these methods, either single-cell sampling or biopsy sampling may be applied. Single-cell sampling means only one cell is chosen for each location and each sample contains 50 single cells from 50 different locations. Biopsy sampling means the 50 cells centered at the sample location are collected. This is repeated for 50 sampling locations, and at each location, one consensus sequence is calculated based on the 50 cells in the biopsy.
Table 2: Parameters range in Latin hypercube sampling. 500 parameters sets were generated within Latin hypercube sampling.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Parameters</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor simulation</td>
<td>Proliferation rate</td>
<td>p0, p1, p2, p3</td>
<td>0.5-1</td>
</tr>
<tr>
<td></td>
<td>Transition probability</td>
<td>t1, t2</td>
<td>0-0.1</td>
</tr>
<tr>
<td></td>
<td>Fitness</td>
<td>f1, f2, f3</td>
<td>1-5 (f0=1)</td>
</tr>
<tr>
<td></td>
<td>Death rate</td>
<td>d0, d1, d2, d3</td>
<td>0-0.5</td>
</tr>
<tr>
<td>Sequence simulation</td>
<td>Passenger mutation accumulation rate</td>
<td>μ1, μ2, μ3</td>
<td>10^{-10} - 2×10^{-9} (μ0=10^{-10})</td>
</tr>
</tbody>
</table>

2.2.2 Different sampling methods of tumor cells

We used 8 different sampling methods, including 4 different sampling locations and 2 different sampling modalities (Figure 6B) to generate 50 samples for one tumor. In single-cell sampling, each sample consisted of one cell. In biopsy sampling (equivalent to in silico “biopsies” or bulk sampling), each sample consisted of 50 adjacent cells and one consensus sequence is calculated from the reads obtained from these 50 cells. A consensus sequence is constructed by assigning the nucleotide found in more than 50% of the relevant aligned reads to each position along the consensus. In random sampling, cells are chosen uniformly at random from the entire sampling area. In core sampling cells are sampled uniformly at random in a disk centered around the tumor’s origin, and of surface equal to half of the whole tumor area. In boundary sampling, cells are sampled from outside the core. Finally, in section sampling the cells were sampled across a line (cross-section) connecting the tumor center to the boundary.
2.2.3 Sequence simulation

Once cell samples are generated and their genealogies have been retrieved, the neutral passenger mutations are superimposed on each lineage, with a mutation rate according to the cell types starting at the base of the genealogy. The mutations are accumulated based on an infinite-sites model (73, 74), which means no mutations will occur twice at the same site. The number of mutations along each branch of the genealogy is a random variable drawn from a Poisson distribution with rate obtained by multiplying the branch length by 30Mb (corresponding to the size of the human exome) and a per-generation mutation rate that varies between $10^{-10}$ to $10^{-9}$ per site per replication cycle (75-77). The mutation model assumes equal nucleotide frequencies, with equal transitions between each pair of nucleotides.

2.2.4 Latin Hypercube sampling of parameter space

To explore the parameter space defined in Table 2, we used Latin Hypercube Sampling (LHS) (78) to generate combinations of parameter values, each chosen from a range of possible values (Table 2). LHS is a stratified Monte Carlo sampling method that is commonly used to explore high-dimensional parameter spaces. It is an efficient sampling strategy, because unlike exhaustive grid-sampling where the number of simulations increases as more parameters are added, LHS allows a user to specify the total number, $c$, of simulations required. Each parameter is sampled over $c$ values (79-81) in combination with other values of other parameters. In this study, we generated 500 combinations of parameters.
2.2.5 **Tree reconstruction and measurements**

We used a distance based method Neighbor-Joining (NJ) (82) with Jukes-Cantor (JC) (83) model of evolution to build phylogenies for our simulated samples. Neighbor-Joining has been shown to be robust when constructing trees (84, 85) and its accuracy is similar to other more time-consuming methods for relatively small data sets (85, 86). This method has also been applied to several tumor phylogenetic reconstruction studies [18,23].

We used the recorded true lineage as the true tree, then measured the symmetric distance (Robinson–Foulds metric (87)) between reconstructed trees and true trees. The symmetric distance was scaled by the maximal possible distance between two trees. We also calculated eight other measurements, including phylogenetic diversity (PD, calculated as the sum of all branch lengths on a tree), variation in branch length (Var), the internal to external branch length ratio (IERatio) and four tree shape statistics: $I_c$, $B_2$, cherries (88) and $A1$ (89), and a test of the correlation between the spatial distance and genetic distance (Mantel’s R (90)) for both reconstructed trees and true trees. Details can be found in supplementary methods. The phylogenies were built with the APE package (91) in R, symmetric distance were calculated with the PHANGORN package (92) in R, Mantel’s test was performed with the VEGAN package (93) in R, and tree statistics were calculated with SIMMONS (94).
2.3 Results

2.3.1 Recovery of the true phylogeny

We compared the symmetric distance between reconstructed trees and true trees. Trees built with single-cell sampling had a mean normalized symmetric distance of 0.33±0.15 while trees built with biopsy sampling had a mean normalized symmetric distance of 0.69±0.19. We found that trees reconstructed with the single-cell sampling method had a significantly smaller sympatric distance to the true lineage compared to biopsy sampling (t-test with p-value < 2.2e-16). This suggests that single-cell sampling can recover trees that are closer to the true phylogeny than those obtained using biopsy (Figure 7). Examples of trees built with different sampling methods are shown in Figure 8. The differences between reconstructed single-cell phylogenies and the true phylogenies are caused by short internal branches, which have insufficient support for any specific topological arrangement of lineages. This effect should be reduced when mutation rates increase. Indeed, we observed a further reduced symmetric difference of 0.09±0.11 (Figure 7). However, it does not improve the reconstruction accuracy with biopsy sampling.
Figure 7: Distribution of scaled symmetric difference between true and reconstructed lineages. Solid lines and dashed lines represent single and biopsy sampling respectively. Different colors represent different sampling location. 10-fold higher mutation rate.
Figure 8: Examples of true and reconstructed lineages with different sampling methods. Blue, yellow and red tips represent Type 1, Type 2 and Type 3 tumor cells respectively. Identical clades are highlighted with boxes of matched shades. Symmetric difference between reconstructed and true lineages are shown below the trees.
2.3.2 Measurements profiles of different sampling methods

Biopsy sampling yielded significantly lower total branch length and branch length variation but higher internal-external branch length ratio compared to true trees, regardless of the sampling location (Table 3). This is not surprising because when consensus sequences of biopsies are used, low-frequency mutations are averaged out, thus the total number of mutations on a phylogeny is lower than one would expect with single-cell sampling. In addition, low frequency mutations are more likely to exist on tips rather than on internal branches, which results in relatively longer internal branches and shorter tip branches. On the other hand, there is no significant difference in total branch length and branch length variation between single-cell sampling and the true tree.

For tree shape measurements, high \textit{Ic}, \textit{B2} and low \textit{cherry} represent highly unbalanced trees. Our results showed that all sampling methods reconstructed more unbalanced trees than the true phylogenies (Table 3) for some tree statistics, (i.e., \textit{PD}, \textit{Var} and \textit{IEratio}). In addition, biopsy sampling tends to produce more unbalanced trees than single-cell sampling (paired t-test, all \textit{p}-values<2.2e-16).

For different sampling locations, Core sampling and Section sampling had higher total branch lengths, branch length variation and internal-external branch length ratio compared to Random and Boundary sampling (true in both single and biopsy sampling, paired t-test, all \textit{p}-value<2.2e-16). This is because Core and Section samples are likely to contain more tumor cells and less normal cells (Figure 9). Since tumor cells have a higher
mutation rate than normal cells, we are likely to observe more mutations in Core and Section samples.

For the correlation between physical distance and genetic distance, a significant Mantel’s R is more frequently obtained with Core and Section sampling than with Random and Boundary sampling for either single or biopsy sampling (Table 4). This is presumably because in our simulation tumor cells had a higher mutation rate than normal cells, thus it is easier to detect the correlation between genetic distance and physical distance in tumor cells. Therefore, the sampling methods that included more tumor cells (Core and Section sampling) were more likely to show a significant Mantel’s R. Whenever Mantel’s R was significant (Empirical significance value<0.01), Section sampling gave the highest correlation. This is because Section sampling contains more tumor cells compared to Random or Boundary sampling and covers a broader physical distance compared to Core sampling, which together give Section sampling the advantage over other methods when showing the overall physical distance and genetic distance correlation.
Table 3: Tree balance statistics comparison of true lineage and reconstructed ones.

<table>
<thead>
<tr>
<th>Stat</th>
<th>TRUE</th>
<th>Random</th>
<th>TRUE</th>
<th>Core</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>0.77</td>
<td>0.77</td>
<td>0.13**</td>
<td>0.80</td>
</tr>
<tr>
<td>Var</td>
<td>7.32E-05</td>
<td>7.34E-05</td>
<td>3.17E-06**</td>
<td>3.22E-04</td>
</tr>
<tr>
<td>IERatio</td>
<td>0.32</td>
<td>0.32</td>
<td>0.57**</td>
<td>2.25</td>
</tr>
<tr>
<td>IC</td>
<td>0.15</td>
<td>0.21**</td>
<td>0.25**</td>
<td>0.12</td>
</tr>
<tr>
<td>B2</td>
<td>4.01</td>
<td>2.91**</td>
<td>2.77**</td>
<td>4.47</td>
</tr>
<tr>
<td>Cheery</td>
<td>15.55</td>
<td>14.70**</td>
<td>13.79**</td>
<td>15.8</td>
</tr>
<tr>
<td>A1</td>
<td>0.63</td>
<td>0.42*</td>
<td>0.32**</td>
<td>0.61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stat</th>
<th>TRUE</th>
<th>Boundary</th>
<th>TRUE</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>0.78</td>
<td>0.78</td>
<td>0.17**</td>
<td>0.78</td>
</tr>
<tr>
<td>Var</td>
<td>9.31E-05</td>
<td>9.31E-05</td>
<td>7.06E-06**</td>
<td>2.03E-04</td>
</tr>
<tr>
<td>IERatio</td>
<td>0.53</td>
<td>0.53</td>
<td>0.84**</td>
<td>1.44</td>
</tr>
<tr>
<td>IC</td>
<td>0.14</td>
<td>0.21**</td>
<td>0.25**</td>
<td>0.22</td>
</tr>
<tr>
<td>B2</td>
<td>4.13</td>
<td>2.93**</td>
<td>2.79**</td>
<td>3.6</td>
</tr>
<tr>
<td>Cheery</td>
<td>16.2</td>
<td>15.21**</td>
<td>14.38**</td>
<td>16.13</td>
</tr>
<tr>
<td>A1</td>
<td>0.77</td>
<td>0.8</td>
<td>0.71</td>
<td>1.2</td>
</tr>
</tbody>
</table>

*p-value <0.01, ** p-value<0.001 (paired t-test between true tree and reconstructed tree)
Table 4: Mantel’s test. The proportion of Mantel’s tests that show a significant relationship between genetic and physical distance and the average Mantel statistics, calculated over 500 simulations across different sampling methods.

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Core</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Single</td>
</tr>
<tr>
<td>% of sig</td>
<td>1.000</td>
<td>0.343</td>
</tr>
<tr>
<td>Avg Mantel</td>
<td>0.474</td>
<td>0.316</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Boundary</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
<td>Single</td>
</tr>
<tr>
<td>% of sig</td>
<td>1.000</td>
<td>0.422</td>
</tr>
<tr>
<td>Avg Mantel</td>
<td>0.410</td>
<td>0.323</td>
</tr>
</tbody>
</table>

Figure 9: Average cell type composition of single-cell samples obtained with different sampling methods. White represents the normal cells and other colors represent different type of tumor cells. Core and Section sampling methods contain significantly lower percentage of normal cells compared to Random and Boundary sampling (paired t-test, all p-value<2.2e-16).
2.4 Discussion

Our results show that reconstructed phylogenies of tumor samples sequencing data are likely to suffer biases resulting from sampling methods. Both sample location and modality affect the reconstructed trees and summary statistics we obtained from the same simulated tumors. Single-cell sampling can provide the best accuracy for phylogeny topology reconstruction and branch length estimation. The four sampling strategies (i.e., Random, Core, Boundary and Section) do not have significant accuracy difference on topology or branch length estimation. On the other hand, biopsy sampling rarely recovers the true tree and gives lower total branch lengths, lower variation of branch lengths, higher internal-external branch length ratios and highly unbalanced trees. Interestingly, phylogenies derived from single-cell sequences also appear to be more unbalanced than the true trees, although to a lesser extent than those derived from biopsy samples.

The accuracy of phylogenetic reconstruction can have downstream effects on our ability to test alternative models of tumor progression. For instance, Townsend and colleagues (65) discussed two biopsy sampling papers (95, 96) in detail and noted that the observed monophyly of metastatic sub-clones does not provide sufficient evidence to support a history of metastatic cascades, i.e., metastasis-to-metastasis spread, because parallel metastases, in which metastases originate from the primary tumor, can also result in a monophyletic pattern when the sub-clone samples in the primary tumor have not been completely sampled. We would further add that the ability to accurately
recover phylogenies also depends on whether we use single-cell samples or biopsy samples.

We also need to consider the composition of tumor samples, which depends on both the sampling location and tumor spatial pattern. In our simulation, a tumor starts to develop from the center, thus core and section sampling will contain more tumor cells than normal cells. And because we assume tumor cells have higher mutation rates than normal cells, these two sampling methods tend to give higher total branch lengths, branch length variation and internal-external branch length ratios compared to random and boundary sampling. Although this may be an over simplified tumor spatial pattern, it indicates that the identity of single cells (tumor or non-tumor) in single-cell sampling or the purity of tumor cells in biopsy sampling should be taken into account when interpreting measurements from samples. If the aim is to characterize tumor cells only, adjustment probably needs to be made when contamination from normal cells is present. A newly developed tumor phylogeny inference method called Treeomics (58) considers the robustness under different tumor purities. This is an appealing approach and one that deserves further exploration.

We showed that a correlation between genetic distance and physical distance could be detected by core single-cell sampling, section single-cell sampling or possibly section biopsy sampling, but not other methods. This suggests that when we compare the genetic diversity of multi-regional samples obtained from different tumors, adjustments based on physical distance may be necessary. For example, the between-
region sub-clonal genetic divergence from multi-region sequencing in tumor has been used to distinguish tumors driven by strong positive sub-clonal selection from neutral or weak case (97). The multi-regional samples used were at least 3cm apart and with purity of 72-96% (97), which may be comparable to the section biopsy sampling and thus adjustments of genetic divergence by physical distance may further improve its classification performance.

In conclusion, we used a simulation model to provide practical suggestions on phylogenetic analysis sampling strategies within tumor, which serves as an initial link between evolutionary models and clinical application. Our model showed that sample location and modality do affect the tree and summary statistics we obtained from the same tumor. Single-cell sampling is the best option for phylogeny topology reconstruction and branch length estimation, while it is still possible to give more unbalanced trees when mutation rate is low. Biopsy sampling rarely recovers the true tree and gives lower total branch lengths, lower variations of branch lengths, higher internal-external branch length ratios and highly unbalanced tree. Furthermore, it is important to consider the purity of samples and physical locations when interpreting these measurements.
3 Inference of sub-clone spatial distribution with phylogenies of phased pseudo-haploids

3.1 Introduction

In cancers, tumor cells that may have originated from one or a few progenitor cells nonetheless accumulate mutations at a sufficiently high rate to become genetically heterogeneous. It is important to understand intra-tumor heterogeneity because of its ability to drive disease progression and the acquisition of therapeutic resistance (5, 98).

As noted in Chapter 1, there is overwhelming evidence of genetic heterogeneity in a variety of neoplasms. However, less focus has been placed on the spatial distribution of genetic heterogeneity. In particular, whereas it has been shown that there is a sub-clonal structure to the genetic variation within tumors, with genomes clustering into discrete haploids, little is known about how these sub-clones are distributed across a tumor.

It is important to understand the distribution of sub-clones because it would be helpful when designing a representative sampling set. If sub-clones are distributed evenly in the tumor, we may detect all sub-clones with one deep sequencing sample obtained from a single location. However, if the sub-clones have strong spatial structures, multiple samples from different locations would be necessary to obtain a representative sample of the genetic heterogeneity in the tumor. When comparing samples from metastases to the primary tumor and building a combined phylogeny,
incomplete sub-clone sampling in the primary tumor may lead to a false conclusion of monoclonal origin of metastases(65).

Spatial information also plays a vital role in explaining the level of heterogeneity. For example, in a multi-regional sequencing study, two bulk samples were taken for each tumor(97), and the genetic divergence between the two samples were used to distinguish strong or weak selection model(97). However, besides the selection pressure, spatial constraints can also shape between-region sub-clonal genetic divergence. Specifically, a more spatially structured population would decrease the within-regional genetic diversity and increase the between-regional genetic divergence(99). Therefore, spatial distribution information is essential for distinguishing alternate evolutionary models.

To resolve spatial distribution patterns in cancer evolution, one possibility is to use single-cell sequencing or a combination of single-cell sequencing and bulk sequencing. For example, single-cell sequencing has been performed in breast cancer(43), liver cancer(100), myeloma(50), bladder cancer(101) and kidney cancer (59). On the other hand, whole genome single-cell sequencing is not yet easily accessible due to the high cost and relatively lower accuracy (32). Another possible method is to employ computational sub-clone de-convolution analysis. Several computer programs including PhyloWGS(57), subcloneSeeker(54), SciClone(53), CLOE(102), LICHeE(103) and Treeomics(58) have been developed to address this issue. Despite the advances in methodology, these methods perform best with relatively small number of sub-clones
and require a high sequencing depth to guarantee accuracy. Another limitation that these methods have is the requirement of location information, so that the resolved sub-clones can be map back to show the spatial distribution patterns. This works well when several locations within a tumor have been sampled and sequenced separately and location information is recorded.

Here we present a new method, which only requires multiple bulk-sequencing genomic data to measure the spatial heterogeneity of sub-clones without the need for location information. It is important to recognize that in the phylogenetic literature, it is inappropriate to build phylogenies using consensus sequences. Nonetheless, in practice, researchers have often gone on to do exactly this by sequencing bulk samples, as noted in Chapters 1 and 2. Therefore, although it would be preferable to use single-cell samples, this chapter is motivated by a commonly used experimental protocol in cancer research. Our method takes advantage of the diploid nature of the human genome. We describe the rationale for the method below, and we show that it allows us to distinguish between tumors with different levels of spatial genomic heterogeneity. We also apply our method to empirical genomic data from liver, breast and colon cancers.

3.2 Methods

3.2.1 Overview

In an ideal sampling design where single cells within a tumor are sequenced individually, we are able to use a phasing algorithm to reconstruct each haploid set of
chromosomes in a diploid genome within a cell. If n cells are sampled, there will be 2n haploid sets of chromosomes. When all 2n haploid sets are analyzed phylogenetically, we expect that cells that are closely related will have homologous chromosomes that are also closely related; hence, we expect the evolutionary tree to consist of two clades, each with n tips, that have identical sub-topologies (Figure 10).
Figure 10: Illustration of sub-clone spatial distribution and phylogeny. Cell lineages within one tumor are shown in (a). Different colors represent different sub-clones. Based on the same tumor lineage (a), there could be different spatial distribution of sub-clones (b). Dotted circles represent locations where multi-regional sequencing samples are taken. A sample is more likely to contain cells from one sub-clone when sub-clones are spatially clustered together, while more likely to contain cells from different sub-clones when they are spatially mixed. In our simulation, we mix different numbers of clades (defined by mixing level) to generate samples representing different sub-clone spatial distribution (c). Then we infer the two haploids for each sample. After that, a tree for each set of haploids is built and we compare the RF distance between them (d).

Now, extrapolate to a sampling design where instead of single cells, we have a bulk sample of multiple cells at each of n locations. If there is very little mixing of cells in
a tumor, most or all cells at each location are likely to be closely related and share a common ancestor not shared with cells from other locations. In this case, it would still be possible to identify a haploid set of chromosomes at each location, where the set is a consensus of the chromosomal sequences. Although it is not guaranteed, one expects that these consensus sequences approximate the chromosomes present in the local progenitor cells. An evolutionary tree of these 2n consensus haploid sets should therefore result in patterns similar to that seen with single-cell haploid sets, i.e., a phylogeny with two clades each of n tips that have (nearly) identical topologies.

In contrast, if there is considerable mixing of cells within a tumor, such that cells that originate from the same progenitor cell are separated in space, then the canonical chromosomes that are reconstructed at each location are no longer derived from cells that share a common progenitor not shared with cells elsewhere. A phylogeny of all 2n consensus haploid sets will no longer necessarily recover a tree with two clades of identical topologies (Figure 10).

The procedure we propose here relies on quantifying the difference between the two sub-topologies of consensus haploid chromosomal sets. The input data are simulated single diploid tumor cells and clades are collapsed to represent sub-clones. We pre-process the data by combining cells from different sub-clones to generate different level of mixed bulk samples. One consensus sequence is calculated for each bulk sample. Once this is done, we apply a phasing algorithm for each sample using
BEAGLE(104). Then we build and compare the phylogenies of each haploid set. We also applied our method to three empirical datasets.

3.2.2 Simulation of tumor cells

We simulated an expanding tumor population of size 1000, 5000 and 10000 with pure birth model(105). Birth rate was set to 0.5. The mutation rate in human was estimated to be around $10^{-8}$ per site per generation (106-108) and mutation rate in tumor is generally considered higher than the baseline mutation rate. The estimated mutation rate in breast cancer is $10^{-6}$ (109) and has high variation across type and across genome(110-112). Therefore, we used a range of mutation rate from $10^{-9}$ per site per generation to $10^{-6}$ per site per generation (Table 6).

3.2.3 Simulation of sequences

To simulate the two haploids for each individual, the ancestor was assigned a homogeneous diploid genome and then point mutations were imposed along the tree branches with and without the assumption of infinite sites (113). A Jukes-Cantor model (114) which assume equal base frequencies and mutation rates was used for nucleotide substitution model.

3.2.4 Sampling and mixing

Once we have the complete lineages and sequences simulated. We generated 10, 20 and 50 samples from the population. Each sample is consisted of 20 cells and these cells come from different numbers (defined by mixing level) of clades. To represent different sub-clone mixing level, we first define sub-clones as a clade of
closely related cells such that each clade has 8-12 tips. Therefore, for population sizes of 1000, 5000 and 10000, there are 100, 500 and 1000 sub-clones respectively. Then to represent a spatially clustered case, each sample is composed of cells from one clade. To represent a mixture case, cells from multiple clades will be combined together to form one sample. The process was illustrated in Figure 10C. We used mixing level of 1, 5, 10 and random. Random mixing means all 20 cells are randomly chosen from the entire population regardless of the clades. This is different from the largest possible mixing level that each cell is guaranteed to come from different clades. All parameters used were shown in Table 5. We ran 100 simulations each for all 1008 combinations of parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>1000, 5000, 10000</td>
</tr>
<tr>
<td>Mutation Rate</td>
<td>10⁻, 5×10⁻, 10⁻, 5×10⁻, 10⁻, 5×10⁻, 10⁻</td>
</tr>
<tr>
<td>Sample size</td>
<td>10, 20, 50</td>
</tr>
<tr>
<td>Mixing level</td>
<td>1, 5, 10, Random</td>
</tr>
<tr>
<td>Switching error rate</td>
<td>0%, 5%, 10%, 20%</td>
</tr>
</tbody>
</table>

3.2.5 Tree reconstruction and comparison

For the multi-regional samples in one tumor, each sample contains multiple cells and each cell has two haploids. We first calculated the consensus haploids for each
sample. Then we separated the two haploids from each sample and built independent phylogenetic trees based on each haploid set using both Neighbor-joining(82) and Maximum Parsimony(115) methods for tree building. Robinson–Foulds distance (RF)(116) was used to measure the distance between each pair of trees. These steps were performed using the APE(91) and Phangorn(92) packages in R.

3.2.6 Haploid phasing switch error

Since the haploid information is inferred rather than sequenced directly, it is reasonable to assume errors exist in the phasing step. Therefore, we tested for the sensitivity of our method regarding errors. We used the definition of switching error used in BEAGLE(104). BEAGLE is reported to have a switch error rate lower than 8% (104). With an infinite sites model, only two possible alleles could be present at each site. We added errors to the haploids on 5%, 10% and 20% of the parsimony informative sites by randomly choosing a parsimony informative site and converting it to the alternative allele.

3.2.7 Compare the sub-clone mixing level in different type of cancer

We also applied our method to three sets of empirical data(15, 117). These data were taken from multi-regional sequences of different types of tumors. We use BEAGLE(104) to estimate the haploids in the samples with default settings. Then we built two neighbor-joining trees based on each set of haploids. The set of haploids was determined by clustering the most similar haploids together.
3.3 Results

3.3.1 Robinson-Foulds distance increases with higher sub-clone mixing level

Using simulations with find that the Robinson-Foulds distance increases with higher mixing level of clones given adequate mutation rate, same population size and sample size (Figure 11). When mutation rate is larger than $10^{-8}$, the RF distance is lowest when each sample comes from only one clade (mixing level = 1) and highest when each sample comes from random clade (mixing level = random), with other mixing levels in between. This suggests that when adequate mutation signals are present, RF distance between two sets of haploids can be used as an indicator of sub-clone spatial mixing level. Low RF distance close to 0 is a strong support for spatially clustered sub-clone. High RF could suggest high mixing level or low mutation load. This is also true with parsimony as the tree-building method.
3.3.2 Adequate mutation rate is required for RF distance to show difference under different sub-clone mixing level

Our results showed that with different population sizes and sample sizes, the separation between different mixing levels becomes more evident with increased mutation rate. This suggests that an adequate number of mutations, which depends on both mutation rate and population size in our simulation, is required to separate different sub-clones. Given the same tumor size, a higher mutation rate means the tumor population can accumulate mutations with higher frequency. Under the same
growth rate, it takes more time to reach a larger size. And given the same mutation rate, more mutations could be accumulated in tumors of larger size. High mutation rate will decrease the overall RF distance, but makes the separation among different mixing levels more evident. Low mutation rate will result in high RF across all mixing level, since not enough signal is present.

For empirical data, it may not be easy to measure the mutation load in the whole tumor. Nevertheless, one possible way to separate the high mixing level and low mutation load case is by comparing them to the random distribution of RF distance. As shown in Figure 11, when mutation rate is as low as $10^{-9}$, the observed RF distances are even higher than the mean of random RF distances. This suggests that higher than random RF distances means that the mutation load may not be enough to distinguish mixing level.

### 3.3.3 Larger sample size can differentiate sub-clone mixing level on finer scale

Larger sample sizes increase the overall RF distance (Figure 11). This is because distributions of RF distance depend on the size of trees. The mean RF distance between random trees increases with tree size. In other words, given the same RF distance, larger sample sizes would support a smaller mixing level. Therefore, we need to consider the sample size when comparing the RF distance in different tumors. It would be best to use the same sample size when comparing mixing level across different tumor or cancers.

Furthermore, larger sample sizes may be able to distinguish mixing on finer scales. Before scaling, RF distance is discrete and the possible values it can take depends
on tree size, \(n\). The number of possible RF values for a tree of size \(n\) is \(n-3\) and the maximum RF distance value is \(2n-6\) (118). For example, the number of possible RF distance values for tree of size 10, 20 and 50 are 7, 17 and 47 respectively. Normalized RF distance is calculated as the ratio between RF distance and the maximum possible value (118, 119). This means after normalizing by the maximum possible value, RF distances of larger sized trees can take values on a finer scale (more possible values on decimal scale). Although sample of size 10 do follow the increasing RF distance for increasing mixing levels, its ability to distinguish mixing levels is limited. Thus, sample sizes larger than 20 are preferable.

3.3.4 Presence of phasing switch error does not affect the increasing trend of RF distance

Haploid phasing inference is not error-free. The program we used in our study, BEAGLE, has a reported switch error rate lower than 8\% (104). To test the sensitivity to error, we introduced 5\%, 10\% and 20\% of error on the parsimony informative sites in the inferred haploids and then reconstructed the trees. Our results showed that randomly introduced errors again keeps the RF distance low when mixing level is low, but tends to increase when mixing level is high. And it still can provide significant separations for mixing levels of 1, 5, 10 and random up to 20\% errors. One example under population size 10000 and sample size 50 is shown in Figure 12. This suggests that we can use this method when phasing inference is not perfect.
3.3.5 Empirical data

We applied our method to one set of colon tumor data (117), one set of liver tumor data (15) and one set of breast tumor data (Figure 13), which were all used multi-regional bulk sequencing of solid tumor.
Figure 13: Phylogenies based on each set of estimated haploids in colon, liver and breast tumor.

For the colon tumor case, we obtained a RF distance of 0.33 with 12 samples including 3 from liver metastasis site. Mutation rate was estimated to be around 1.7 per...
MB in colon rectum cancer(120). Considering the phasing error, this value is on the lower end of RF distance distribution, which suggests spatial clustering of sub-clones exists in this case. However, given the relatively small number of samples, it is difficult to estimate the exact mixing level.

For the liver cancer case, we obtained a RF distance of 0.55 with 23 samples and mutation rate was estimated to be 1.5 per MB in liver tumor(120). Consider the phasing error, this is lower than mean of random RF distance and around middle of observed values from our previous simulations, which may suggest there are multiple sub-clones present and some spatial structure may exist.

For the breast cancer case, we obtained a RF distance of 0.79 with 32 samples. Mutation rate in breast cancer was estimated to 1.0 per Mb in breast cancer(120). If we assume the phasing inference is perfect, which is less likely, since the RF distance is larger than the mean of random RF distances for tree of size 32, it would suggest not enough variants are present for resolving the sub-clone distribution. If we take the errors during phasing into consideration, which increases RF distance in higher mixing level cases, this would suggest a high mixing level of sub-clones in this breast tumor.

3.4 Discussion

We inferred the sub-clone mixing level to be low in colon tumor, and high for liver and breast cancer. Colon cancer follows a distinct pattern featuring progression from single crypt lesion to invasive adenocarcinomas(121). The data set we analyzed also contain 3 samples from a metastatic site. It’s growth pattern and the sampling
strategy all contribute to the physical separation of cells and sub-clones. Therefore, it’s reasonable to observe lower RF distances. Another study of multi-regional sequencing in colon cancer also showed that sub-clones were generally separated in geographically correlated ways(39). For liver cancer, the original paper used 35 SNV found by single-cell sequencing to define 20 sub-clones and then genotyped 286 bulk samples to map the sub-clone distribution(15). Despite samples straddling two clones, they made the general conclusion that genealogically, separate clones were observed to be segregated, revealing limited cell movement within solid tumors. This agrees with the intermediate RF distance we obtained from this data set. Navin’s group have performed single-cell sequencing for breast tumor(42). In one case of 100 cells from 6 sectors of one tumor, they identified 5 subpopulations. None of the subpopulations is unique to any sector and all sectors contains multiple subpopulations. This also suggests a possible high spatial mixing level in breast tumor.

Note that our phasing method returns phased pseudo-haploids rather than estimates of the actual haploids. Because each sample contains multiple cells, there are probably more than two haploids that exist in each sample. However, we only estimate two pseudo-haploids from each. We may consider them as the average (or “canonical”) haploids from each sample. Nevertheless, based on our results, they can still provide information about spatial distribution of sub-clones in multiple samples. Another limitation of our method is that it does not consider structural variations, including those resulting from mitotic recombination. These processes will further complicate the
haploid estimation and increase the estimation error. However, there is evidence suggesting that structural variations on chromosomes occur early in the tumor development, and may already fixed in the tumor population (43). Based on our phasing error analysis, strong spatial clustering pattern has stable low RF signature even under the influence of high error rate. Thus, a small RF distance would still likely suggest spatial clustering of sub-clones.

Our method can provide suggestions for sequencing strategy for different types of cancer. To study the complexity of heterogeneity in tumor, sub-clone spatial distribution should be considered when choosing sequencing methods. Evidence has been shown that different type of cancer has different clonality levels (23, 122). For tumors with more mixed clone distribution, ultra-deep sequencing followed by a de-convolutional method would be more favorable than multi-locus sequencing. Because in this case, each multi-locus sample is equally representative of the population. On the other hand, for a spatially clustered tumor, one ultra-deep sequencing sample is likely to miss sub-clones that exist at a distant location. Therefore, multi-locus samples would be necessary for this type of tumor.

We presented a new method that can measure the tumor clonal mixing level with multi-regional bulk sequencing data and is free of sample location information requirement. To achieve better resolution, larger sample size (>=20) is suggested. Small RF distance (close to 0) would strongly suggest spatially clustered sub-clones. However, caution needs to be taken when interpreting larger RF distances since these may result
from high mixing levels, low mutation load or high phasing error rate. One principle is if
the RF distance is larger than the mean RF distance between random trees of the same
size, it is likely to suggest that not enough mutations are present. Furthermore, to
compare across multiple tumor cases, same number of samples is suggested for fair
comparison.
4 Method validation and evolutionary analysis of breast cancer, using multi-regional bulk sequencing data

4.1 Introduction

In this chapter, we aim to use the methods developed in previous chapters to investigate clonal evolution in two multi-regional bulk sequencing cases in breast cancer. We used phylogeny and sub-clone de-convolution to study the evolutionary relationship between invasive cells and tumor cells within the primary tumor, and between invasive lymph nodes and primary tumor in breast cancer. We also investigated the sub-clonal distribution patterns within the primary tumor with the method we developed in Chapter 3.

4.1.1 Breast cancer

Breast cancer is the most prevalent cancer in women both in the developed and the developing world. According to WHO, nearly 1.7 million new cases were diagnosed in 2012(123). In US, there were over 250,000 estimated new cases in 2017 making up 30% of all cancer cases(124). Breast cancer survival rates vary greatly worldwide, ranging from 80% or over in North America, Sweden and Japan to around 60% in middle-income countries and below 40% in low-income countries(125).

The diagnosis of breast cancer usually involves various imaging methods including mammograms, ultrasounds and MRI to distinguish normal and diseased tissue. Other procedures including blood tests, bone scans, computerized tomography (CT) scans or positron emission tomography (PET) scans may also be used. Despite advances in imaging, tissue biopsy sampling is still considered the definitive method of diagnosis.
Breast cancer stages range from 0 to IV with 0 (carcinoma in situ) indicating a cancer that is noninvasive or contained within the milk ducts, and Stage IV indicating a cancer that has spread to other areas of the body. The most widely used method of staging breast cancer is the TNM system, where TNM stands for tumor size, lymph node status and metastases respectively. Starting in 2018, new measures of tumor grade, estrogen receptor status, progesterone receptor status and HER2 status were added to the TNM system. Each stage has specific combinations of TNM, grades and receptors status.

4.1.2 Ductal carcinoma in situ (DCIS) and Invasive ductal carcinoma (IDC)

Ductal carcinoma in situ (DCIS) is generally an early, non-invasive form of breast cancer. However, it can be found along or with invasive breast cancer. In DCIS, the abnormal cells are contained in the milk ducts (canals that carry milk from the lobules to the nipple openings during breastfeeding, Figure 14). It is the most commonly diagnosed precursor of breast cancer and one of the most common pre-cancers across all tissues with approximately 60,000 diagnosed each year in the U.S.

Invasive ductal carcinoma (IDC) means the abnormal cancer cells that began forming in the milk ducts have spread beyond the ducts into other parts of the breast tissue (Figure 14). At this point, it may be able to spread (metastasize) to other parts of the body through the lymph system and bloodstream. It is the most common type in breast cancer, making up nearly 70-80% of all breast cancer diagnoses.
4.1.3 Metastasis in breast cancer

More than 90% of the solid tumor-related mortality corresponds with metastasis (131). However, metastasis shows diverse behaviors and organ dissemination patterns in different types of cancer. Pancreatic cancer and small-cell lung cancer are often present with widespread metastases to multiple organs. Glioblastoma usually remains locally invasive within the central neural system. In breast cancer, metastases can be undetectable at the time of initial diagnose and keep latent for many years. The primary metastasis sites for breast cancer are bone and lung, followed by liver and brain (132).

Approximately 10–15% of patients with breast cancer have an aggressive disease and develop distant metastases within 3 years after the initial detection of the primary tumor (133). The spread of tumor cells can happen through the lymphatic system or bloodstream. Approximately one third of women who have breast cancer with no lymph node involvement at time of diagnosis will develop distal metastases (133), about one-third of patients with breast tumors that have spread to the lymph nodes remain free of
distant metastases 10 years after local therapy (134, 135).

It is estimated that about 40-50 percent of DCIS cases may progress to invasive breast cancer if left untreated (136). However, currently which cases of DCIS will progress to invasive breast cancer is unclear, thus almost all cases of DCIS are recommended to surgery treatment with a certain percentage of overtreatment. At the other extreme, some patients initially treated for DCIS later develop widely disseminated metastatic disease, indicating that some cases are being undertreated.

4.1.4 Models of cancer clonal evolution

One of the earliest cancer evolution models is a step-wise process and is supported by cancer incident data at different time points from patient populations (137, 138). This became the standard model of cancer development and is often referred to as the “selective sweep” model. In this model, several “driver” mutations need to occur and undergo positive selection to fixation, replacing the whole population.

The selective sweep model (Figure 15A) was originally thought of as a linear model where mutations are acquired sequentially and rapidly take over the whole neoplastic population. In 1976, Nowell proposed a clonal model which allows branching evolution of clones (6). However, it’s not until the advances in sequencing technologies, that evidence for branching patterns (Figure 15B) has started to accumulate. The use of multi-biopsy strategies and genome-wide analysis provides irrefutable evidence of
multiple co-existing clones and intra-tumor heterogeneity(8, 13, 16, 22, 139).

The Big Bang model(35) adopts the branching model and proposes that once the precursor lineage of a cancer has accumulated the necessary driver mutations, the tumor grows as a single clonal expansion under neutral growth (Figure 16B). Different from the selection sweep model, the frequencies of clones within a tumor is a consequence of the time since appearance rather than its relative fitness. It would also produce a higher level of heterogeneity than the selective sweep model. Genomic profiling of 349 individual glands from 15 colorectal tumors showed an absence of selective sweeps, uniformly high intra-tumor heterogeneity and sub-clonal mixing in
distinct regions, consistent with the predictions of Big Bang model(38).

Despite the detection of neutral growth, selection is likely to occur at some stage during clonal evolution in order for a neoplasm to grow within normal tissues. However, quantitative measurement of the strength of positive selection remains a major challenge with current evolutionary experiments(33). To explain the observation of both population cancer incidence data in the selective sweep model and neutral growth within tumor, a new model of punctuated equilibrium was proposed(33), in which tumor populations undergo a series of different phases: punctuation, gradualism and stasis (described in Chapter 1).

Figure 16: Models of cancer clonal evolution. (cited from Graham and Weight(33))
4.1.5 Models of invasive cell evolution

The fact that some, but not all, histologic diagnoses currently categorized as “pre-cancers” evolve into cancer immediately leads to perhaps the most fundamental clinical and biological questions: which pre-cancers are most likely to evolve into cancers? To answer this question, it’s necessary to understand how invasive cells arise from the primary tumor.

To describe the possible relationship among normal cells, tumor cells and invasive tumor cells, Navin’s group proposed three different models (Figure 17) considering the progression of DCIS to invasive ductal carcinoma(IDC): (1) independent evolution; (2) evolutionary bottlenecks; and (3) multi-clonal invasion(140). In independent evolution, different normal cells give rise to DCIS and invasive component separately. In the evolutionary bottleneck model, one normal cell give rise to both DCIS and invasive component. And DCIS and invasive component have a sequential development pattern, with all invasive components arise from one of the DCIS cells. In multi-clonal invasion, similar to evolutionary bottleneck, DCIS and invasive components all arise form one normal cell. However, multiple DCIS clones can escape and cause invasion.
4.1.6 Evidences for models of invasive cell evolution

A multi-regional targeted sequencing study of 303 samples form 50 invasive breast cancer patients showed branched patterns of clone development and in two cases the pre-chemotherapy and chemotherapy-resistant metastases all arose from the sub-clones in primary tumor(38). The finding of an early origin of metastasizing sub-clones supports the multi-clonal invasion model. A single-cell copy number profiling study of 200 cells for 2 breast tumor cases supports a punctuated clonal evolution model(141). A recent single-cell CNV study of ten DCIS and IDC patients also showed support for multi-clonal invasion model within the primary tumor(44). They used Topographic Single-cell Sequencing, which can preserve the location information, on
1293 single cells from 10 patients. Their results support the hypothesis that multiple clones escaped from the ducts and co-migrated into the adjacent tissues to establish invasive carcinomas. They also found all clones were detected in both the in situ and invasive regions, while some sub-clones were more restricted to the ducts or invasive regions.

Supports for different models also exist in other type of cancers. A sequencing study of seven pancreatic cancer metastases show branched patterns indicative of tumor evolution(142). A whole-genome bulk sequencing of 26 metastases from four patients with pancreatic cancer showed that metathesis to lung and liver also followed a multi-clone invasion pattern(58, 143). However, another multi-regional exome sequencing study of renal carcinomas and associated metastatic sites showed a bottleneck pattern of metastasis(12). These all suggest much can be learned about cancer development by the application of evolutionary models (5, 144-146).

More evidence is needed to understand the evolutionary processes operating on of invasive tumor cells and the extent to which such processes differ between and within different type of cancers. To our knowledge, no study so far has tracked the invasive components both within the primary tumor and the lymph nodes in breast cancer. Therefore, we performed multi-regional exome sequencing on two DCIS cases with invasive components. And one of the cases has invasive tumor cells present on multiple lymph nodes. Using phylogenetic analysis, we can infer their evolutionary history and investigate their genetic and spatial origin in the primary tumor.
4.2 Methods

4.2.1 Multi-regional bulk whole-exome sequencing in two DCIS cases with invasive component

We analyzed two DCIS cases with an invasive component collected in the laboratory of Dr. Shelley Hwang’s group at Duke School of Medical. Case 1 is estrogen positive, progesterone positive, and HER-2 negative (ER+/PR+/HER2-). Case 2 is positive on all three receptors (ER+/PR+/HER2+). No evidence of recurrence was found in both cases. The physical measurements of the biopsy samples were 8cm x 7cm x 3cm and 6.9cm x 4.7cm x 2.3cm respectively.

For each case, the tumor was divided into a series of slices and each slice was further divided to a different number of blocks. For Case 1, 30 samples were taken in total, including one benign sample from lymph node and 29 samples from one tumor. For Case 2, 26 samples were taken in total, including three invasive samples from lymph nodes, 22 from one tumor and one benign sample from another tumor of the same patient. Details of sampling locations are illustrated in Figure 18a. For each sampled block, if both DCIS and invasive cells are present, they are isolated by microdissection and sequenced separately (Figure 18b).
Figure 18: Tumor dissection and sampling locations. a. For each case, tumor was divided into a series of slices and then each slice was further divided to a different number of blocks. White block represents no existence of any tumor cell. Grey block represents the existence of DCIS cells. Orange stripes represent the existence of invasive tumor cells. Orange dots labels the blocks invasive samples were taken. Block with a ID number represents DCIS samples were taken. b. For blocks that contain both DCIS and invasive cells, they were separated with microdissection (stained tissue slides contributed by Dr. Hwang’s group).
4.2.2 Whole exome sequencing

Sequencing was performed at the McDonnell Genome Institute at Washington University School of Medicine in St. Louis. For library construction, 20 ng of DNA was used to generate Illumina sequencing libraries as dual-indexed libraries using unique bar-code identifiers and the Swift Biosciences library kit. In addition to exomes, we also included an 83-genes panel for breast cancer. Hybrid capture with the 83-gene panel was performed by combining the library pools with the IDT probe set for 72 hours at 47°C followed by stringent washing to remove spuriously hybridized library fragments. Enriched library fragments are eluted following isolation with streptavidin-coated magnetic beads and amplified prior to sequencing. DNA sequencing is performed on an Illumina HiSeq 2500 1T instrument. We obtained ≥80% of the targeted regions at ≥30X coverage.

4.2.3 Variant calling

Raw reads were aligned with BWA MEM(147) and prepared following GATK best practice guide(148), including duplicates removal, realignment around indels and recalibration on base quality score. Variant calling was performed by GATK HaplotypeCaller, Platypus(149) and Mutect(150). Mutect was shown to be one of the two callers that achieved the highest sensitivity among five somatic mutation specific callers in one benchmark study (151). However, these somatic mutation callers usually take one normal sample and only one matched tumor sample to call the somatic mutations. But our data requires joint-calling of multiple tumor samples. Therefore,
besides the somatic mutation caller MuTect, we also took advantage of germline callers GATK and Platypus which can call variants on multiple samples simultaneously.

We use the union set of all three callers and performed a post filtering step to further reduce the data to a high-confidence set of mutations. The combination of multiple pipelines is suggested by multiple variant-calling benchmark studies (152, 153). Union of the variants were then filtered with all the following filters: 1. only one ALT allele present, 2. at least one tumor genotype calling pass all filters and differ from normal control, 3. mutation is shared by at least two tumor samples.

Identified somatic mutations were then annotated with Oncotator (154) and pathway enrichment analysis was performed with Bioconductor(155).

4.2.4 Phylogeny

We selected somatic SNVs to be used for tree building. Additional filters for somatic SNVs are (all filters should be passed):

1. no missing information (zero read depth) for all samples
2. minimum variant frequency 0.05
3. minimum number of total reads containing alternative allele 5
4. minimum number of forward reads supporting alternative allele 2
5. minimum number of reverse reads supporting alternative allele 2
6. no reads containing alternative allele in normal control

These sites were concatenated into one sequence for each sample.

Subsequently, one Neighbor-Joining tree(82) with a Jukes-Cantor(114) model and one
maximum parsimony tree were built. To assess the reliability of trees, 1000 bootstrap replicates were performed for each tree. Tree building and bootstrapping was done with FastME 2.0(156), APE(91) and PHANGORN(157) packages in R(158).

4.2.5 Haploid estimation and sub-clone spatial distribution estimation

To infer the spatial clustering level of sub-clones, we compared the trees built with each phased pseudo-haploid set using the method developed in Chapter 3. Pseudo-haploids were estimated from genotype data with Beagle’s phasing algorithm(104). To minimize the effect of copy number variation, somatic mutations of loss of heterozygosity (LOH) were excluded from haploid estimation. One Neighbor-joining(82) tree was built for each set of haploids. Robinson–Foulds(RF) metric (87) was used to compare the topologies of these two trees.

4.2.6 Sub-clone estimation

To validate our estimation of the spatial distribution of sub-clones we also employed another tool Treeomics(58) which can reconstruct the phylogeny of metastases and map sub-clones to their anatomic locations. For sub-clone estimation, we used the list of exonic mutations on potential driver genes found in our dataset.

4.2.7 Correlation between genetic distance and physical distance

Linear regression was used to test if a correlation exists between genetic distance and physical distance. Sample 3 and Sample 1 are chosen as the reference sample for Case 1 and Case 2 respectively since they are located on the most superior end of each tumor. All distances were calculated with respect to the reference sample.
Physical distances were calculated as Euclidian distance of block coordinates and scaled by physical dimensions. Genetic distances are calculated as number of somatic SNV difference on potential driver genes. For Case 2, samples from the node are excluded from analysis since the physical distance is unmeasured. Linear regression was performed in R(158).

Mantel’s test(90), a method that specifically designed for testing correlation between two distance matrixes was also used and performed with VEGAN package(159) in R.

4.3 Results

4.3.1 Phylogeny of SNVs supports multiple origins of invasive cells

Based on the phylogenetic tree (Figure 19), both cases showed polyphyletic patterns, supporting multi-clone development of invasive cells. Based on the reconstructed phylogenetic tree, for Case 1, three invasive samples in the primary tumor arise from two different sub-clones of tumor cells. And for Case 2, ten invasive samples from the primary tumor arise from 3-4 different sub-clones of tumor cells.

This multi-clonal pattern is further supported by sub-clone analysis (Figure 20). In Case 1, there are 9 sub-clones estimated in total and invasive cells arise from sub-clone 2, 3 and 8. In Case 2, there are 2 sub-clones estimated in total and invasive cells arise from both of them.
Figure 19: Phylogenetic trees. Both Neighbor Joining and maximum parsimony methods are used for phylogenetic tree building. Trees are rooted on normal control sample. Each tip on the tree represents one sample. Invasive sample labels are highlighted in orange. Branch length is scaled to the evolutionary distance used in phylogenetic tree analysis. For parsimony tree, distance is the number of character-state change and for Neighbor Joining tree the distance is based on Jukes-Cantor(114) model of DNA substitution. Bootstrap (1000 replicates) support value ranging from 0 to 1 are shown on the node. Larger bootstrap value represents better support and values above 0.5 are colored in red. Branches which located differently between trees build with different methods are colored in blue.
4.3.2 Invasive cells in primary tumor and lymph nodes may follow different origin models

In Case 2, all three invasive samples from the lymph nodes arise from the same invasive sample clade on the phylogenetic tree (monophyletic) (Figure 19). Again, this is supported by the sub-clone analysis that all the three invasive samples on lymph nodes arise from a single sub-clone 2 (Figure 20). This suggest that although invasive cells within the primary tumor follow a multi-clonal model, the invasive cells that spread to lymph node may not. However, we cannot rule out the possibility that the observation
of this bottleneck pattern for lymph node invasion is due to incomplete sampling of the primary tumor.

4.3.3 Timing of invasive cells and mutations

Phylogenies of both cases showed that the invasion not necessarily all arise from the most recent clades (Figure 19). They could arise from earlier clade as well, which supports an early seeding of invasive cells. This is also validated by the sub-clone analysis (Figure 20). For Case 1, invasive cells can arise from the early sub-clones 2 and 3. For Case 2, invasive cells can arise from the first funding sub-clone 1.

The phylogenetic tree for Case 2 shows relatively long terminal branches and short internal branches. This pattern suggest large amount of mutations may occur in a short burst of time, supporting the Big Bang and punctuated equilibrium model of tumor progression (33, 35).

4.3.4 Sub-clone and spatial distribution

The large RF distance between trees built with each haploid set (Figure 21) suggests strong spatial clustering of sub-clones may not exist at least in our two samples. This is also supported by the sub-clone deconvolution analysis. For example, in Case 2, sub-clone 2 gave rise to multiple samples taken across the tumor (Figure 20), which suggests this clone is widespread in the primary tumor.
Figure 21: Trees built with estimated haploid sets. Haploids were estimated from genotype data with Beagle 4.1. Then one NJ tree was built for each separated haploid set.
4.3.5 Significant correlation between genetic distance and physical distance

Linear regression between genetic distance and physical distance (x, y, z coordinates based on block map) showed significant correlation for both cases (Figure 22). Another correlation test on the two matrices (Mantel test) showed significant correlation in Case 1 but not in Case 2. This may because that in Case 2, there are nine pairs of DCIS and invasive cells sampled from the same block, thus sharing the same physical location coordinates. Therefore, the physical location information for Case 2 is less distinct due to major overlapping of DCIS and invasive samples.

Figure 22: Linear regression of physical distance and genetic distance. Physical distances are calculated as Euclidian distance of block coordinates between each tumor sample and reference sample. Genetic distances are calculated as number of genotype difference between each tumor sample and reference sample. Both distances are normalized in the figure. Sample 3 and Sample 1 are chosen as reference sample for Case 1 and Case 2 respectively since they locate on the most superior end of each tumor. For Case 2, samples from the node are excluded form analysis since the physical distance is unmeasured. Color of sample points represents different slices the sample located on. Linear regression line is shown with grey area represents the 95% confidence interval. Both regression has p-value<0.5, suggesting a significant correlation between physical distance and genetic distance.

4.3.6 Mutations in potential driver genes

We identified mutations in potential driver genes and their distribution across all samples is shown in Figure 23. We find that except for SETD2 in Case 2, the majority of driver gene mutations are sub-clonal (synapomorphies), meaning they exist in some but
not shared across all samples. The ATP-binding cassette (ABC) transporter pathway, which contributes to drug resistance via ATP-dependent drug efflux pumps (160), has enriched mutations in both cases. There is no driver mutation identified that is unique to invasive cells.

![Figure 23: Heat map of somatic mutations on potential driver genes](image)

### 4.4 Discussion

We observed branching patterns for tumor cells in the primary tumor as well as for potential driver mutations (Figure 20). The short internal branch lengths (Figure 20 Case 2) may indicate a rapid increase in the numbers of cancer cells, supporting the Big Bang and punctuated equilibrium models (33, 35). This is consistent with other multi-regional studies (38) and single-cell studies (44, 141) in breast cancer. However, this contradicts studies in other cancer type (12). These findings suggest different cancers may undertake different evolutionary models.
For the invasive cells in primary tumor, our results support the multi-clonal model of invasion in primary tumor, which agrees with a single-cell study in breast tumor(44). However, another single-cell study of CNV in breast tumor suggested monoclonal origin of invasive cells(42). The difference may be explained by the different data types used. Previous studies suggest that chromosomal rearrangements occur early in cancer evolution and are followed by stable clonal expansions to form the tumor mass, resulting in highly similar CNV profiles and diverse somatic mutation profiles(43). However, for invasive cells on lymph nodes, we did not find evidence for multi-clonal model.

Our study also suggests that invasive cells may arise at an early stage in primary tumors. This agrees with another single-cell study in breast cancer(43) and an multi-regional study in breast cancer(38). These patterns have significant clinical relevance. If metastatic disease can arise from a very early branch of the phylogenetic tree, it may not carry the mutation a targeted treatment would aim for. In this scenario, targeted treatment would not help prevent disease relapse(38).

We observed spatial mixing of sub-clones, which agrees with the Topographic Single-cell Sequencing study of breast tumors(44). They also showed that majority of the clones defined by both copy number or genotype mutations were detected in both the in situ and invasive regions. The difference is although their spatial resolution of single cells is higher than our block samples, it’s restricted to a 2-dimensional slice while our sampling is across the 3-dimensional tumor. However, we can still detect a significant
correlation between physical distance and genetic distance with mutations that define sub-clones. This is interesting, and may indicate either that sub-clone spatial structure and mixing co-exist in tumors, or that correlations between genetic and physical distances, and our method of phasing pseudo-haploids can conflict under certain conditions. Although small RF distance would be strongly supported for spatially clustered sub-clones, large RF distances may result from lack of adequate mutation signals or high phasing error.

There are a few limitations of the present study. The number of cases we investigated is two. We did not sample all the tumor blocks in our cases, which may lead to sampling bias. We focused only on exome sequencing and did not cover other possible variations on copy number, structural variations, variations in non-coding regions, expression or epigenetic variations. These could be future directions for a more comprehensive study of breast tumor evolution.
5 Validation and interpretation of single-cell sequencing of lung cancer

5.1 Introduction

In this chapter, we aim to investigate one non-small cell lung cancer case with single-cell sequencing. Specifically, we are interested in the intra-tumor heterogeneity, evolutionary processes in early-stage non-small cell lung cancer (NSCLC) and their potential association with clinical outcome. Currently we have obtained 12 single cells in one case. We will present the preliminary results and interpretations in this chapter.

5.1.1 Lung cancer

Lung cancer is the second most common cancer and the leading cause of cancer related mortality (161). It’s projected to cause over 150,000 deaths in 2018, equal to the mortality of the next four most deadly cancers combined (breast, prostate, colon, and pancreas)(161). Lung cancer symptoms tend to not appear at early stages, thus most patients diagnosed with lung cancer today have advanced disease (40% are stage IV, 30% are stage III), and the current the 5-year survival rate is only 16% (162).

Traditional diagnosis of lung cancer was based on pathology and categorized into three types according to the size of affected cells. The most common type (80%-85%) is non-small cell lung cancer (NSCLC), which include adenocarcinomas, squamous cell (epidermoid) carcinoma, large cell (undifferentiated) carcinoma and others(163). The second type (10%-15%) is small cell lung cancer (SCLC) (163). Both NSCLC and SCLC may occur at the same time. SCLC is often very aggressive and tumor grows quicker than NSCLC. The third type (5%) is lung carcinoid tumors made up of neuroendocrine cells;
this particular type of lung cancer is uncommon and tends to grow slower than other
types of lung cancers (163).

Low-dose CT scans are recommended for high risk populations for screening.

However, although this strategy can detect slightly more lung cancers, smaller tumors,
and more stage I tumors, the detection of a larger number of early-stage cancers is not
accompanied by a reduction in the number of advanced lung cancers or a reduction in
lung cancer deaths (164).

5.1.2 Non-small cell lung cancer (NSCLC)

Non-small cell lung cancer is an umbrella term for several types of lung cancers
include squamous cell carcinoma, adenocarcinoma and large cell carcinoma (Figure 24).

Adenocarcinoma arises from glandular cells of bronchial mucosa. Squamous cell cancer
arises from the modified bronchial epithelial cells (thin, flat cells that line certain
organs). Large cell neuroendocrine carcinoma is a malignant epithelial tumor, which is
comprised of large polygonal cells that do not show any obvious evidence of histological
differentiation (165).
5.1.3 Prognostic markers

The assessment of prognosis is important for the selection of appropriate treatments in each individual case. Main prognostic factors in patients with lung cancer are tumor stage and performance status (166). Main staging system of lung cancer is the TNM system, where TNM stands for tumor size, lymph node status and metastases respectively. However, for patients with clinical stage I lung cancer, they have at best a 60% 5-year survival rate following surgical resection, suggesting that at least 40% of all stage I patients have undetectable metastatic disease at the time of presentation (167, 168). Therefore, there is an urgent need for a more accurate molecular classification of
lung cancer based on biological attributes of cancer rather than on gross anatomic findings.

Potential markers including protein markers, DNA methylation, gene expression, microRNA and images analysis have all been explored. However, large-scale analysis of the lung adenocarcinoma genome, transcriptome, and methylome also revealed that integrated subtypes characterized by idiosyncratic combinations of molecular alterations underscore the heterogeneity of this disease. This suggests that any one molecular biomarker may not be sufficient for a good prognosis marker. Understanding heterogeneity and its origin may provide new insights for lung cancer prognosis.

5.1.4 Intra-tumor heterogeneity in lung cancer

Genetic intra-tumor heterogeneity in lung cancer has been observed in several studies. One recent study used 327 multi-region whole-exome sequencing on 100 early-stage NSCLC tumors to investigate the evolutionary histories, obtain a census of clonal and sub-clonal events, and assess the relationship between intra-tumor heterogeneity and recurrence-free survival. They observed parallel evolution at the level of chromosome-arms but not on mutational level, and detected positive selection for late, but not early, mutations. For prognosis, they found that patients who had tumors with a high proportion of sub-clonal copy-number alterations were at significantly higher risk for recurrence or death than those with a low proportion, but there was no significant association between the proportion of sub-clonal mutations and relapse-free survival.
Another multi-regional sequencing study of 25 spatially distinct regions from seven operable NSCLCs found evidence of branched evolution, with driver mutations arising before and after sub-clonal diversification(71). Another multi-region whole-exome sequencing (WES) study on 11 localized lung adenocarcinomas found on average, 76% of all mutations and 20 out of 21 known cancer gene mutations were identified in all regions of individual tumors, suggesting that single-region sequencing may be adequate to identify the majority of known cancer gene mutations in localized lung adenocarcinomas(15). With a median follow-up of 21 months after surgery, three patients have relapsed, and all three patients had significantly larger fractions of sub-clonal mutations in their primary tumors than patients without relapse(15). These data indicate that a larger sub-clonal mutation fraction may be associated with increased likelihood of postsurgical relapse in patients with localized lung adenocarcinomas(15).

These multi-regional sequencing studies provided valuable information for intra-tumor heterogeneity, but they have limitations on the resolution and number of samples per tumor. Furthermore, based on our simulation in Chapter 2, phylogenies built with bulk sequencing may have systematic biases and would benefit from validation with other methods. Single-cell sequencing would be ideal for phylogeny building. To the best of our knowledge, there is one single-cell copy number study in lung cancer on circulating tumor cells(CTCs)(178). However, they focused on detecting characteristic cancer-associated single-nucleotide variations and insertions/deletions in exomes of CTCs and generating highly reproducible results. Therefore, an intra-tumor
single-cell study would be ideal for investigating the early evolution history of primary tumor.

5.2 Methods

5.2.1 Data set

The case we investigated is a Stage I NSCLC patient, who had two separate adenocarcinomas and a squamous cell carcinoma. The squamous cell recurred. We performed single-cell whole-exome sequencing for 12 of the recurrent squamous cells and one single cell from a cell-line. Freshly resected lung tumor was dissociated into a single cell suspension using a Tumor Dissociation kit (Miltenyi Biotec GmbH, Germany) according the manufacturer’s instructions. The suspension was cleared of macroscopic debris by filtering through a 70 µm nylon screen and lymphocytes, fibroblasts, and endothelial cells were removed with a Miltenyi Tumor Cell Isolation kit. Recovered tumor cells were re-suspended in Bambanker (Wako Chemicals, Richmond, VA) cell freezing medium and frozen at -80°C.

After thawing, cells were sorted individually into the wells of a 96-well PCR plate. Sorted cells were gated as alive/CD45<sup>neg</sup>. The plate was covered with plate sealing tape and frozen at -80°C until DNA amplification. Sequencing was done using Illumina’s MiSeq, a next-generation sequencing platform. We reached mean depth of x33.

5.2.2 Mutation calling

Raw reads were aligned to reference genome GRCh37 with BWA MEM(147) and prepared following GATK best practice guide(148), including duplicates removal,
realignment around indels and recalibration on base quality score. Variants were called separately for each sample with HaplotypeCaller in GVCF mode and then jointly genotyped together.

5.2.3 Phylogenetic analysis

For phylogenetic analysis, we selected 21285 variant sites with no missing information for all samples (read coverage>0). These sites were concatenated to one sequence for each sample. We used Smart Model Selection (SMS)(179) with Akaike Information Criterion to select the Generalized Time Reversible (180) model as the best fitted nucleotide substitution model for our data. Then we used PhyML (181) to build maximum-likelihood trees with this model. An Approximate Likelihood-Ratio Test (aLRT) (182) was used to provide the branch support. We used mid-point rooting for the trees. Tree shape statistics Ic, B2, cherries (88) and A1 (89) were calculated with SIMMONS(88), as was done in Chapter 2.

5.2.4 Mutation profiling

Identified somatic mutations were annotated with Oncotator (154). Pathway enrichment analysis was performed with Bioconductor(155). Mutation tree was built with non-synonymous mutations on potential driver genes with SCITE(61).

5.2.5 Estimation of effective population size dynamics

To investigate the past population dynamics, we used skyline plot to extract historical effective population size information from shape of the phylogeny. We used BEAST(183) for skyline plot, parameters include: site model of GTR, clock model of
relaxed exponential, tree model of Coalescent Bayesian skyline and 5,000,000 MCMC steps.

5.3 Results

5.3.1 Phylogeny

The phylogenetic trees built with each replicate sets are the same (Figure 25) and they show generally good support (aLRT >0.75) except for two internal nodes.

The tree shape measurements for this tree showed low Ic, B2 and high cherry values (Table 6), representing a more balanced tree. In chapter 2, we showed that single-cell sequencing tends to generate more balanced trees than biopsy bulk sequencing. Here at least for this specific tree, it follows our prediction in chapter 2.
Figure 25: Maximum-Likelihood tree of 12 single cells in one patient and 1 cell-line control. E3B is the cell-line control and the other tips each represent one single cell from the same patient. The branch length is a joint estimation of mutation rate and time. Branch support is the Approximate Likelihood-Ratio Test score (aLRT) ranging from 0-1, with higher number represents better support.(Note that the letters and numbers of sample are merely identification IDs, same letter/number does not suggest any relationship between the two samples)

Table 6: Tree statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>PD</em></td>
<td>0.66</td>
</tr>
<tr>
<td><em>Var</em></td>
<td>0.001</td>
</tr>
<tr>
<td><em>I</em>Eratio</td>
<td>0.37</td>
</tr>
<tr>
<td><em>Ic</em></td>
<td>0.29</td>
</tr>
<tr>
<td><em>B2</em></td>
<td>2.62</td>
</tr>
<tr>
<td><em>Cherry</em></td>
<td>5</td>
</tr>
<tr>
<td><em>A1</em></td>
<td>6.25</td>
</tr>
</tbody>
</table>
5.3.2 Mutation profiling

We identified 21,285 variant sites (both germline and somatic mutations), of which 14,010 were sub-clonal (exists in some but not all cells), thus likely to be somatic mutations. Exonic mutations exist in 43 potential driver genes. One of the driver gene mutation on FRG1B is clonal while others are all sub-clonal. Mutation enriched pathways includes EGFR tyrosine kinase inhibitor resistance, endocrine resistance, Rap1 signaling pathway, central carbon metabolism in cancer, hepatocellular carcinoma and renal cell carcinoma (p<0.0001). We used 54 non-synonymous mutations on potential driver genes to build a mutation tree. The mutation tree showed parallel branching pattern for RELN, KMT2C, MECOM, NCOR1, EP300 genes (Figure 26b).
Figure 26: Heatmap of potential driver mutations and mutation tree. a. Heatmap of exonic mutations on potential driver genes. b. Mutation tree estimated by SCITE. Each node represents a non-synonymous mutation on a potential driver gene. Branching pattern of mutated genes are highlighted in different colors. Each cell sample is attached as grey circles.
5.3.3 Effective population size

Our estimation shows small effective population size and a decreasing trend of tumor cells. Broadly, effective population size represents the number of individuals in a population who would contribute offspring to the next generation. A small effective population size may suggest the existence of cancer stem cells (184), since only a small amount of cells contribute to the next generation. Strong selection and fixation on cells carrying advantageous mutations may also contribute to the small effective population size (185). An increasing trend of effective population size suggests tumor population is expanding, consistent with the growth of recurrent tumor in this patient.

Figure 27: Estimation of effective population size. a. Skyline plot shows a decreasing population size. Y-axis is effective population size. X-axis represents amount of change inferred from SNVs. The black line is the median estimate of the estimated effective population size. Blue area is the upper and the lower estimates of 95% interval. b. Tracer plot of MCMC chain. Grey period is the burn-in. c. Marginal distribution of posterior and prior.
5.4 Discussion

We showed that we were able to build a phylogeny for single-cell sequencing in lung cancer. The tree presented characteristics consistent with our simulation in Chapter 2. We observed a branching pattern of tumor cells but did not find evidence for punctuated evolution in our sample. Given the limited sampling, of course, this may not be generalizable to other lung cancer cases. More data is needed for further validation and interpretation of the clonal history.

There is some evidence showing a correlation with mutation burden and the clinical outcome of patients\(15\). Summary statistics of phylogenetic trees like phylogenetic diversity, internal-external branch length ratio and topology measurements may open up more opportunities for the development of reliable prognostic markers. We will further test this hypothesis when more data is available.

Single-cell sequencing can identify far more sub-clonal mutations than multi-regional bulk sequencing. A multi-regional study of 100 patient with Stage I lung cancer identified sub-clonal mutations ranging from 2 to 2310 \(177\). However, we identified 14,010 sub-clonal mutations with 4855 mutations on exome regions. Previous studies showed driver mutations in \textit{EGFR, MET} were almost always clonal\(177\). But mutations on these two genes in our samples are sub-clonal. We observed parallel branching patterns for multiple driver genes, for example \textit{EP300 and RELN} which is consistent with other studies\(15, 71\). One limitation of our study is that due to the lack of a normal control sample from the patient, we cannot separate clonal somatic mutations that are
shared by all cancer cells from germline mutations that also present in all cells, which means we may miss some driver mutations that occur earlier in the tumor development.

We also show that it is possible to estimate the dynamics of effective population size from single cell phylogenies. Effective population size may play an important role in our understanding of tumor evolution. A small effective population size may suggest the existence of cancer stem cells, such as in colon cancer(184). It may also suggest strong selecting is acting on cells with advantages mutations. Their fixation can reduce the effective population size. However, for populations with small effective population sizes, genetic drift could be more predominant than selection, which means a “clonal” mutation we observed in tumor population may not necessarily carry fitness advantages but neutral or even deleterious. This may contribute to cancer heterogeneity and the reason why it is hard to cure. Furthermore, we estimated an expanding effective population size, which fits the recurrent growth of tumor. Although we only tested 1 case, it suggests using effective population size dynamics as a prognostic marker for recurrent/non-recurrent patient may be feasible. Our analysis shows the potential usage of effective population size estimation, and it may help predict the prognosis and clinical outcome of tumor.
6 Conclusions

In this dissertation, I explored how existing phylogenetic analysis can be applied to decipher tumor evolution and developed novel method addressing sub-clone spatial distribution within tumor. These model, method and analysis opens up new opportunities for integrating evolutionary biology and cancer research, and may lead to better understanding, predicting and controlling of this disease.

In Chapter 2, using in silico simulations, I took the first step of finding the optimal sampling design for tumor sequencing studies, where the goal is to use phylogenetic to study the evolution of cancers. My results indicate that sampling design is important for accurate estimation of phylogeny and appropriate methods should be chosen accordingly. However, there are several limitations in my model. It only considers 2-dimensional cell growth with a limited number of cells. Tumor also interacts with its microenvironment including connective tissue, blood vessels, and inflammatory cell. The empirical sampling procedure is likely to be more complicated than in my simulation. However, my modelling framework is flexible and could be expanded to include more scenarios.

In Chapter 3, I developed a new method for inferring the sub-clone spatial distribution within tumor, which can be applied to bulk sequencing data when no location information is available. Although it does not consider structural variation and may be inclusive when large RF distances are observed, it does not rely on location information, and thus may be used as a validation for sub-clone de-convolution analysis.
In Chapter 4, I demonstrated the use of phylogenetic analysis and applied the method developed in Chapter 3 with multi-regional bulk sequencing data in breast cancer. It shows, for the first time, that invasive cells in situ and invasive cells on lymph nodes may have different evolutionary histories. In Chapter 5, I showed the application of phylogenetics to single-cell sequencing in lung cancer. This is the first single cell study of primary tumor in non-small cell lung cancer. I demonstrated that it is feasible to construct the evolutionary dynamics of population size, based on these data.

The use of phylogenetic analysis in cancer studies is growing rapidly. It can be used to answer old questions: can we distinguish potential “driver” mutations that provide a fitness advantage from passenger mutations, and can we develop new prognostic markers based on the shape and properties of phylogenetic trees? It may also be possible to ask, in the future if not presently, how the evolutionary history of tumor clones or sub-clones are shaped by conditions within the host or the external environment.

However, tumor phylogenetic analysis has not yet achieved the robustness needed for clinical applications. One problem with using established phylogenetic methods is that these do not provide models or algorithms that incorporate some tumor-specific mechanisms such as chromothriipsis, in which thousands of clustered chromosomal rearrangements occur in a single event. Another problem is that phylogenetic analyses have only been applied to a limited number of isolated cases and
how well an observed pattern supports an underlying hypothesis of cancer progression, e.g., linear vs branched, monoclonal vs polyclonal, is unclear.

I believe that much can be done to address these issues. Simulation studies are ideal for developing principles on how to apply phylogenetic methods on tumors and generate reproducible results with some measure of statistical confidence. For future development, agent-based modeling has the advantage of flexibility, and this is important for incorporating complex tumor cell behavior. The model may take advantage of agent-based simulation framework on GPUs such as the Flame GPU(186) to enable three-dimensional large-scale modeling of tumor.

As shown in Chapter 2, phylogenetic analysis is best use with single-cell sequencing (SCS) data. SCS has also developed rapidly in the recent five years and has already revolutionized our understanding of cancer evolution. Although most of the pioneering studies focused on technology development and validation, the field has started to crack the complex evolutionary processes in cancer. Understanding the origin of metastasis and therapy resistance would provide new insight into the control and treatment of cancer.

SCS may enable us to establish a better correspondence between phenotypic characters and genetic variation. Current cancer clinical staging and classification is largely depended on pathological or morphological characters of tumor cells. However, these characters alone are not enough to explain the variation in treatment response or disease development (133-135, 167, 168). One major obstacle has been the genetic
heterogeneity between and within different cancers and individual patients. SCS may help resolve the genetic heterogeneity on the single cell level, thus has the potential to better connect the phenotype and genotype of cancer cells. SCS may also help us unravel the complexities of cancer evolution dynamics when combined with more sophisticated sampling designs such spatial and temporal sampling.

To summarize, phylogeny has just begun to show its potential to reveal the cancer evolutionary process. There is considerable room for the development of new phylogenetic methods and tools. Simulation studies would play an important role in these developments, since it provides the underlying truth which allows evaluation of these estimation methods. Combined with other advanced biotechnologies such as SCS, the application of these methods in clinical settings may have a major impact on the control of cancer and patient clinical outcomes.
References


86. Takahashi K, Nei M. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood
when a large number of sequences are used. Molecular Biology and Evolution. 2000;17(8):1251-8.


134. Rosen PP, Groshen S, Saigo PE, Kinne DW, Hellman S. Pathological prognostic factors in stage I (T1N0M0) and stage II (T1N1M0) breast carcinoma: a study of 644 patients with median follow-up of 18 years. Journal of Clinical Oncology. 1989;7(9):1239-51.


Biography

Yuantong Ding was born May 20th, 1990 in Jiangsu, China. She received Bachelor of Science degree in Biology in 2012 from Fudan University, Shanghai. After college, she attended graduate school at Duke University and received her Ph.D. in Biology in 2018. While pursuing her graduate degree, she received the following fellowships: Duke University Graduate School Summer Research Fellowship, The Ray J. Tysor Summer Graduate Fellowship, Graduate Travel Fellowship. She has been the recipient of grant funding in Evolutionary Medicine from the Triangle Center for Evolutionary Medicine (TriCEM)/National Evolutionary Synthesis Center (NESCent).