Characterizing cancer genome evolution in metastasis

Ariana Melanie Ivana Huebner

PhD supervisors: Nicholas McGranahan, Charles Swanton, Javier Herrero

A dissertation submitted for the degree of Doctor of Philosophy

University College London
September 2022
Declaration

I, Ariana Huebner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

September 13, 2022

Ariana Huebner
Abstract

Metastatic disease is responsible for the majority of cancer related deaths. Exploring metastatic patterns in an evolutionary framework is crucial to predicting and treating metastases. However, previous studies have mainly relied on archival tissue obtained from early-stage disease and less commonly included samples from recurrent or metastatic disease. In this thesis, genomics data is used to elucidate the tumor’s evolutionary trajectory and to better understand the processes involved in metastatic development.

To better understand the phylogenetic relationships between primary and metastatic samples, especially in extensively sampled multi-region cohorts, a novel method to perform mutation clustering and phylogenetic reconstruction was developed.

Using paired primary metastatic data, the genomic features characterizing clones seeding metastases are explored. Definitions for patterns of metastatic dissemination from the primary tumor are presented and explored within the cohort. Additionally, selection patterns are explored in the seeding clone within the primary tumor compared with selection in other primary specific clones as well as with non-metastasizing tumors.

The relative timing of metastatic divergence is defined and explored in paired primary metastatic data. The importance of extensive sampling of the primary tumor to correctly classify timing of divergence is highlighted. Additionally, a cohort of extensively sampled metastases is presented and leveraged to further characterize the metastatic process. Furthermore, by applying a method to infer migration histories, instances of metastasis-to-metastasis seeding are highlighted.

To understand the process of metastatic evolution in late-stage disease, a cohort of 24 patients with pancreatic ductal adenocarcinoma was processed and analyzed. Using whole exome sequencing of cfDNA samples, the evolution of tumor progression and dynamics of mutations and copy number alterations over time are investigated.

Understanding the processes involved in metastatic dissemination can guide therapeutic management of metastatic tumors as well as lead to the identification of new potential targets for therapy.
Impact statement

Metastatic disease is the leading cause of cancer-related deaths worldwide. The main challenge in managing the disease is treatment resistance due to emergence of a resistant subclone through selective pressures acting on a heterogeneous tumor. Understanding the heterogeneity of not only the primary tumor, but also any arising metastases, as well as better characterization of the processes leading to metastatic dissemination could improve treatment strategies. Over the last few years, the heterogeneity within the primary tumor has been explored, including genomic drivers arising during tumor development. However, the effect of drivers on the overall fitness as well as metastatic potential of cells within the primary tumor remains unclear. My research has focused on characterizing the metastatic process from early to late-stage disease by integrating primary and metastatic samples to investigate tumor evolution and selection over time. Understanding which cells from the primary tumor can lead to metastasis and characterizing recurring patterns, such as selection of specific events, could help inform cancer treatments and lead to new potential targets for therapy. Furthermore, by characterizing which patients’ tumors will likely recur compared to those that do not have metastatic potential, patient stratification for treatment could be improved leading to better patient outcomes.
Acknowledgements

Firstly, I would like to thank my supervisors Dr Nicholas McGranahan, Professor Charles Swanton and Professor Javier Herrero for their support and guidance over the past few years. Particularly, I would like to thank Nicholas McGranahan for his supervision and infectious enthusiasm for the work. It has been an honor to work at the intersections of the Cancer Genome Evolution Research Group and the Cancer Evolution and Genome Instability Laboratory and to witness all the incredible science and researchers who have become friends – I could not have imagined a better PhD experience.

I would also like to thank all of my colleagues across the two groups, past and present, for all their help and input throughout the years. If I were to list all of them and their contributions it would fill this whole thesis, so I will only mention a few, but know that I will always be grateful for the warm and welcoming work environment and exciting discussions about science. In particular Michelle Dietzen, who started this journey with me, and whom I am privileged to call my friend. She has been a constant throughout the PhD and has shared my ups and downs with me along with many hours of discussions about our research. I would also like to thank Maise Al-Bakir, my co-lead on the primary metastasis project, for the incredible collaboration which has shaped the last few years of my work. I am also very grateful to Mariam Jamal-Hanjani for giving me the opportunity to be a part of the PEACE project and for her mentorship and advice throughout.

I would also like to acknowledge the patients and their families, without whom none of this research would have been possible.

Last, but certainly not least, I would like to thank my friends and family for their constant support throughout the years.
# Table of Contents

**Abstract** ........................................................................................................................................... 5

**Impact statement** .............................................................................................................................. 7

**Acknowledgements** ............................................................................................................................ 9

**Table of Contents** ............................................................................................................................... 11

**Table of Figures** ................................................................................................................................. 15

**Abbreviations** ..................................................................................................................................... 19

**Chapter 1  Introduction** ....................................................................................................................... 21

1.1 Tumor heterogeneity ......................................................................................................................... 21
  1.1.1 Genomic variation ....................................................................................................................... 22
  1.1.2 Other key drivers of diversity .................................................................................................... 23
  1.1.3 Implications of heterogeneity .................................................................................................... 24

1.2 Tumor development as an evolutionary process ............................................................................... 25
  1.2.1 Selection .................................................................................................................................. 26
  1.2.2 Neutral evolution ...................................................................................................................... 27
  1.2.3 Signatures of selection .............................................................................................................. 28
  1.2.4 Different tumor growth models ............................................................................................... 28
  1.2.5 Methods to measure selection ................................................................................................. 29
  1.2.6 Region based phylogenetic reconstruction .............................................................................. 32
  1.2.7 Clone based phylogenetic reconstruction ............................................................................... 34
  1.2.8 Sampling bias during clonal reconstruction ......................................................................... 35

1.3 Evolutionary trajectories in metastasis ........................................................................................... 36
  1.3.1 Timing metastatic divergence ................................................................................................. 36
  1.3.2 Modes of metastatic dissemination ......................................................................................... 39
  1.3.3 Metastasis-to-metastasis seeding ......................................................................................... 41
  1.3.4 Metastasis driver events ........................................................................................................... 42

1.4 Tracking tumor evolution over time using ctDNA ......................................................................... 43

**Chapter 2  Data and Methods** ............................................................................................................ 47

2.1 Data .................................................................................................................................................. 47
  2.1.1 TRACERx421 paired primary metastatic cohort ..................................................................... 47
  2.1.2 PEACE cohort ......................................................................................................................... 48
2.1.3 Pancreatic ductal adenocarcinoma cohort ........................................ 49

2.2 Methods ............................................................................................. 49
  2.2.1 Processing of TRACERx and PEACE cohorts ............................... 49
  2.2.2 Processing of pancreatic cohort ....................................................... 56
  2.2.3 Mutational signature deconvolution .............................................. 57
  2.2.4 Quantifying selection using dN/dS ................................................ 57

Chapter 3  TRACERx clustering and tree building methods ....................... 61

  3.1 Introduction ....................................................................................... 61
  3.2 Main restrictions on tree building .................................................... 61
  3.3 Overview of existing methods .......................................................... 65
  3.4 Previous implementation in TRACERx100 ....................................... 65
  3.5 Motivation for updating method ....................................................... 66
  3.6 Updated approach .......................................................................... 67
    3.6.1 Mutation clustering ................................................................. 67
    3.6.2 Phylogenetic tree building ....................................................... 70
    3.6.3 Classifying clonality and presence of individual clusters in tumor regions ..................................................... 72
  3.7 Benchmarking and validating approach ......................................... 73
    3.7.1 Simulation framework for tumor evolution ............................... 73
    3.7.2 Benchmarking different clustering and tree building approaches ................................................................. 75
  3.8 Conclusions ...................................................................................... 78

Chapter 4  Characterizing the seeding clone in metastasis .................... 79

  4.1 Introduction ....................................................................................... 79
  4.2 Modes of metastatic dissemination ............................................... 80
    4.2.1 Defining dissemination patterns from the primary tumor based on shared clonality 80
    4.2.2 Metastatic dissemination patterns in TRACERx paired primary metastasis cohort 83
    4.2.3 Sampling bias due to number of metastasis regions sampled ....................................................... 87
    4.2.4 Determining the seeding clone .................................................. 88
    4.2.5 Comparison of dissemination patterns to MACHINA .................. 89
    4.2.6 Comparison of seeding clones to MACHINA ............................ 92
  4.3 Quantifying selection in the seeding clone ...................................... 98
    4.3.1 Cluster level: Seeding vs primary specific mutations .................. 99
    4.3.2 Paired analysis: Enriched vs depleted ...................................... 108
4.4 Conclusions ...........................................................................................................111

Chapter 5 Characterizing the metastatic process ..............................................115

5.1 Introduction .........................................................................................................115

5.2 Timing of metastatic divergence .......................................................................116
  5.2.1 Defining timing of divergence .................................................................116
  5.2.2 Characterizing early/late diverging metastasis .......................................118
  5.2.3 Sampling bias due to number of primary tumors sequenced .................122
  5.2.4 Timing divergence using mutational signatures .......................................124

5.3 Overview of PEACE cohort ..............................................................................125

5.4 Inferring migration histories ............................................................................133
  5.4.1 Inferring clone proportions .....................................................................134
  5.4.2 Implementing and running MACHINA ..................................................136
  5.4.3 Quantifying occurrences of metastasis-to-metastasis seeding in TRACERx 136
  5.4.4 Quantifying occurrences of metastasis-to-metastasis seeding in PEACE 140
  5.4.5 Timing metastatic migrations ..................................................................144
  5.4.6 Characterizing uncertainty in migration patterns ...................................146

5.5 Conclusions .......................................................................................................149

Chapter 6 Evolution of metastatic disease in pancreatic ductal adenocarcinoma ......................................................................................................................151

6.1 Introduction .......................................................................................................151

6.2 Cohort overview ...............................................................................................152
  6.2.1 Timelines ..................................................................................................152
  6.2.2 Genomic overview ...................................................................................153

6.3 Comparing BRCA versus wildtype patients ..................................................161
  6.3.1 Mutation level comparison of BRCA versus wildtype ..........................161
  6.3.2 Copy number level comparison of BRCA versus wildtype ....................163

6.4 Validation of BRCA vs WT in TCGA .............................................................165
  6.4.1 TCGA overview .......................................................................................166
  6.4.2 Mutation level comparison of TCGA BRCA GL vs WT ..........................166

6.5 Exploring cfDNA samples ..............................................................................168
  6.5.1 High purity cfDNA samples .................................................................169
  6.5.2 Setting filters for calling mutations in cfDNA samples .......................170
  6.5.3 cfDNA samples with signature of unknown origin ..............................172
6.6 Evolution of pancreatic tumors ................................................................. 176
  6.6.1 Adapted CCF estimate and validation .................................................. 177
  6.6.2 Determining copy number aberrations through phasing .................... 179
  6.6.3 Examples of evolution over time within pancreatic ductal adenocarcinoma .... 181
  6.6.4 Patient PAN103 .................................................................................... 182
  6.6.5 Patient PAN113 .................................................................................... 184
  6.6.6 Patient PANVH2 .................................................................................... 187

6.7 Conclusions .............................................................................................. 189

Chapter 7 Discussion .................................................................................... 193
  7.1 Inferring dissemination patterns from the primary tumor and investigating
     the seeding clones ....................................................................................... 193
    7.1.1 Novel clustering and tree building approach in large-scale multi-region
             patients ............................................................................................... 193
    7.1.2 Dissemination patterns from the primary tumor .................................. 194
    7.1.3 The seeding clones present selection compared to non-seeding clones .... 196
  7.2 Characterizing the metastatic process and metastasis-to-metastasis
     seeding ....................................................................................................... 197
    7.2.1 Timing of metastatic divergencence .................................................. 197
    7.2.2 The PEACE cohort allows for better characterization of the metastatic
             process ............................................................................................... 198
    7.2.3 Metastases can seed further sites of disease ....................................... 199
  7.3 Ongoing evolution and heterogeneity in late-stage disease ...................... 200
  7.4 Conclusions .............................................................................................. 201

References ..................................................................................................... 203
# Table of Figures

| Figure 1-1 | Overview of key drivers of diversity. | 23 |
| Figure 1-2 | Overview of the evolutionary process in cancer. | 26 |
| Figure 1-3 | Cartoon visualizing the dN/dS method. | 29 |
| Figure 1-4 | Schematic showing timing of divergence. | 37 |
| Figure 1-5 | Schematic showing metastatic divergence. | 40 |
| Figure 3-1 | Cartoon highlighting the pigeonhole principle applied in tree building. | 62 |
| Figure 3-2 | Cartoon illustrating the crossing rule applied in tree building. | 64 |
| Figure 3-3 | Comparison of dissemination patterns using the previous clustering and updated clustering approaches. | 67 |
| Figure 3-4 | Schematic of updated clustering approach. | 69 |
| Figure 3-5 | Example of output from the tree building method. | 71 |
| Figure 3-6 | Overview of the number of simulations that completed, failed, or timed out for the different methods. | 75 |
| Figure 3-7 | Mutation clustering ARI for the different methods. | 76 |
| Figure 3-8 | Mutation presence precision for the different methods. | 77 |
| Figure 4-1 | Schematic of metastatic dissemination from the primary tumor for individual metastatic samples. | 81 |
| Figure 4-2 | Overview of modes of metastatic dissemination from the primary tumor. | 83 |
| Figure 4-3 | Dissemination patterns per metastatic sample. | 84 |
| Figure 4-4 | Overview of dissemination patterns in TRACERx paired primary metastasis cohort. | 85 |
| Figure 4-5 | Dissemination patterns for each metastatic sample grouped by tumor. | 86 |
| Figure 4-6 | Number of metastatic regions sampled influences definition of dissemination patterns. | 88 |
| Figure 4-7 | Comparison of dissemination patterns between TRACERx and MACHINA. | 91 |
| Figure 4-8 | Comparison of seeding clones between TRACERx and MACHINA. | 93 |
| Figure 4-9 | Example phylogenetic tree of tumor where TRACERx and MACHINA seeding clones differ. | 94 |
| Figure 4-10 | Example MACHINA tree output of tumor where TRACERx and MACHINA seeding clones differ. | 95 |
| Figure 4-11 | Comparison of primary seeding and single source seeding hypotheses of MACHINA in defining seeding clones. | 97 |
Figure 4-12 Cluster overview ................................................................. 99
Figure 4-13 Comparison of mean CCF of the primary tumor for seeding and primary specific clusters. ................................................................. 100
Figure 4-14 Comparison of proportion of primary regions which harbor the seeding and primary specific clusters. ................................................................. 101
Figure 4-15 dN/dS comparison of mutations from seeding and primary specific clusters. ......................................................................................... 103
Figure 4-16 dN/dS comparison of lung cancer genes for subclonal mutations only ... 104
Figure 4-17 Gene level selection comparison between seeding and primary specific clusters. ......................................................................................... 105
Figure 4-18 Gene level dN/dS analysis in LUAD ........................................ 107
Figure 4-19 Gene level dN/dS analysis in LUSC ........................................ 108
Figure 4-20 Schematic of definitions used in paired analysis ..................... 109
Figure 4-21 Paired gene-level analysis ..................................................... 110
Figure 5-1 Schematic of timing of divergence relative to the last clonal sweep. .... 118
Figure 5-2 Timing of divergence .............................................................. 119
Figure 5-3 Timing of divergence split by histology .................................. 120
Figure 5-4 Timing of divergence split by metastasis type ......................... 121
Figure 5-5 Percentage of primary-clonal mutations absent in metastases .... 121
Figure 5-6 Under-sampling the primary tumor can lead to incorrect classification of timing of divergence ................................................................. 123
Figure 5-7 Example of timing divergence using mutational signatures .......... 125
Figure 5-8 PEACE sample overview ....................................................... 126
Figure 5-9 Overview of tissue types sampled within PEACE .................... 127
Figure 5-10 Genomic overview of PEACE cohort .................................... 129
Figure 5-11 Overview of all PEACE trees ............................................... 131
Figure 5-12 Timepoint trees of PEA013 with nodes colored that are present at primary surgery, recurrence, and tissue harvest ........................................ 132
Figure 5-13 Timing of divergence in PEACE .......................................... 133
Figure 5-14 Overview of metastasis-to-metastasis seeding in TRACERx ....... 137
Figure 5-15 Metastasis-to-metastasis seeding in CRUK0029 ..................... 138
Figure 5-16 Metastasis-to-metastasis seeding in CRUK0242 ..................... 138
Figure 5-17 Metastasis-to-metastasis seeding in CRUK0484 ..................... 139
Figure 5-18 Metastasis-to-metastasis seeding in CRUK0090 ..................... 140
Figure 5-19 Migration graph and body map of PEA141 showing reseeding of metastases. .................................................................................................................................................................................................................................................................................................140
Figure 5-20 Overview of different tissue types seeding further metastases. .................141
Figure 5-21 Alluvial plot highlighting source and targets of different migration patterns. .................................................................................................................................................................................................................................................................................................143
Figure 5-22 Clinical and imaging timeline of PEA013. ......................................................144
Figure 5-23 Migration graph and body map of PEA013 highlighting metastasis-to- metastasis seeding.................................................................145
Figure 5-24 First possible migration pattern inferred by MACHINA for PEA043............147
Figure 5-25 Second possible migration pattern inferred by MACHINA for PEA043....147
Figure 5-26 Third possible migration pattern inferred by MACHINA for PEA043.......148
Figure 5-27 Fourth possible migration pattern inferred by MACHINA for PEA043....148
Figure 6-1 Timelines of patients from diagnosis and sample collection to death..............153
Figure 6-2 Genomic overview of cohort. ...........................................................................156
Figure 6-3 Overview of mutational signature proportions for the cohort separated by clonality. .................................................................................................................................................................................................................................................................................................158
Figure 6-4 Overview of copy number events per sample for a set of genes of interest. .................................................................................................................................................................................................................................................................................................160
Figure 6-5 Comparison of mutation burden in patients with and without BRCA alterations. .................................................................................................................................................................................................................................................................................................162
Figure 6-6 Comparison of proportion and mutation counts of SBS3 in patients with and without BRCA alterations. .................................................................................................................................................................................................................................................................................................163
Figure 6-7 Comparison of wGII scores...........................................................................164
Figure 6-8 Comparison of proportion of LOH.................................................................165
Figure 6-9 Overview of pancreatic ductal adenocarcinoma TCGA cohort used to validate findings. .................................................................................................................................................................................................................................................................................................166
Figure 6-10 Comparison of mutation burden in TCGA..................................................167
Figure 6-11 Comparison of SBS3 in TCGA.................................................................168
Figure 6-12 Example copy number profiles....................................................................169
Figure 6-13 Overview of effect of different filters on mutation load and false positive rate. .................................................................................................................................................................................................................................................................................................171
Figure 6-14 Number of cfDNA samples retained for each set of filters. .......................172
Figure 6-15 Heatmap confirming all samples were derived from the same patient. .....173
Figure 6-16 Cosine similarities of the mutational profile and all SBS signatures featured on COSMIC. .................................................................................................................................................................................................................................................................................................174
Figure 6-17 Overview of cfDNA samples with signature of unknown origin. ..........175
Figure 6-18 Transcriptional strand bias of cfDNA mutation profiles. .........................176
Figure 6-19 Overview of samples and tumor content over time. ..........................177
Figure 6-20 Validation of adapted CCF calculation for three samples of one patient. 179
Figure 6-21 Evolution over time of pancreatic tumors. .......................................182
Figure 6-22 Timeline of PAN103 showing treatment overview as well as sample collection..............................................................................................................183
Figure 6-23 Phylogenetic tree of PAN103. ..............................................................184
Figure 6-24 Clinical timeline of PAN113. ..............................................................184
Figure 6-25 Phylogenetic tree of PAN113. ..............................................................185
Figure 6-26 Cancer cell fraction of clusters in PAN113 over time. .........................186
Figure 6-27 Copy number profiles of metastasis and three cfDNA samples of patient PAN113. ...........................................................................................................187
Figure 6-28 PANVH2 ctDNA positivity and response to treatment over time. ............187
Figure 6-29 B-allele frequency plots of PANVH2 as computed by ASCAT. ..............188
Figure 6-30 Phased copy number profiles of PANVH2 showing the germline, the PDX sample, and three cfDNA samples. .................................................................189
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOBEC</td>
<td>apolipoprotein B mRNA editing enzyme catalytic polypeptide-like</td>
</tr>
<tr>
<td>ARI</td>
<td>adjusted rand index</td>
</tr>
<tr>
<td>BAF</td>
<td>B-allele frequency</td>
</tr>
<tr>
<td>CCF</td>
<td>cancer cell fraction</td>
</tr>
<tr>
<td>cfDNA</td>
<td>cell-free DNA</td>
</tr>
<tr>
<td>CIN</td>
<td>chromosomal instability</td>
</tr>
<tr>
<td>ctDNA</td>
<td>circulating tumor DNA</td>
</tr>
<tr>
<td>DBS</td>
<td>double base substitution</td>
</tr>
<tr>
<td>GL</td>
<td>germline</td>
</tr>
<tr>
<td>HLA</td>
<td>human leukocyte antigen</td>
</tr>
<tr>
<td>HR</td>
<td>homologous recombination</td>
</tr>
<tr>
<td>INDEL</td>
<td>small insertion/deletion</td>
</tr>
<tr>
<td>ITH</td>
<td>intratumor heterogeneity</td>
</tr>
<tr>
<td>LN</td>
<td>lymph node</td>
</tr>
<tr>
<td>logR</td>
<td>log-ratio</td>
</tr>
<tr>
<td>LOH</td>
<td>loss of heterozygosity</td>
</tr>
<tr>
<td>LUAD</td>
<td>lung adenocarcinoma</td>
</tr>
<tr>
<td>LUSC</td>
<td>lung squamous cell carcinoma</td>
</tr>
<tr>
<td>MRCA</td>
<td>most recent common ancestor</td>
</tr>
<tr>
<td>MRD</td>
<td>minimal residual disease</td>
</tr>
<tr>
<td>MSAI</td>
<td>mirrored subclonal allelic imbalance</td>
</tr>
<tr>
<td>NSCLC</td>
<td>non-small cell lung cancer</td>
</tr>
<tr>
<td>PDAC</td>
<td>pancreatic ductal adenocarcinoma</td>
</tr>
<tr>
<td>PDX</td>
<td>patient-derived xenograft</td>
</tr>
<tr>
<td>SBS</td>
<td>single base substitution</td>
</tr>
<tr>
<td>SCNA</td>
<td>somatic copy number alteration</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>single nucleotide variant</td>
</tr>
<tr>
<td>STAS</td>
<td>spread through airspace</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TMB</td>
<td>tumor mutation burden</td>
</tr>
<tr>
<td>VAF</td>
<td>variant allele frequency</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>WES</td>
<td>whole exome sequencing</td>
</tr>
<tr>
<td>WGD</td>
<td>whole genome doubling</td>
</tr>
<tr>
<td>wGII</td>
<td>weighted genomic instability index</td>
</tr>
<tr>
<td>WGS</td>
<td>whole genome sequencing</td>
</tr>
<tr>
<td>WT</td>
<td>wildtype</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

Cancer is a group of over 100 diseases where cells grow and spread throughout the body (What Is Cancer?, 2021; Cancer, 2022). It represents the leading cause of deaths worldwide, with metastatic disease accounting for approximately 90% of cancer-related deaths (Chaffer and Weinberg, 2011; Seyfried and Huysentruyt, 2013).

Primary lung cancer is the most commonly diagnosed cancer and leading cause of cancer-related deaths worldwide, with about 18% of cancer-related deaths attributed to lung cancer (Bray et al., 2018). In the United Kingdom, lung cancer is the third most common cancer, however it is the most common cause of cancer-related death (Lung Cancer Statistics, no date). Pancreatic cancer is the 10th most common cancer in the United Kingdom and represents the 5th most common cause of cancer-related death with only 5% of patients surviving more than 10 years (Pancreatic Cancer Statistics, no date). In both tumor types, mortality rates are still high and especially in pancreatic cancer no significant improvement has been observed over the past 50 years. Although overall lung cancer mortality rates have decreased by a third in the past 50 years, lung cancer-related mortality in females have increased by three-quarters. This highlights that although some progress has been made and a shift towards earlier detection can be observed, a more extensive understanding especially of the metastatic process is necessary.

In this thesis, I will be characterizing the metastatic process and exploring the role of intratumor heterogeneity and selection on disease recurrence, as well as investigating ongoing evolution within metastatic disease.

1.1 Tumor heterogeneity

Variations across tumors are observed both within tumors of different tissue types as well as between individuals with the same tumor type. Tumors across different organs will vary in their genomic landscape as well as their prognosis and response to therapy. Therefore, the site of origin will dictate treatment decisions. However, even when the cancer originates in the same tissue, variable response to treatments can be observed across individuals (Weigelt and Reis-Filho, 2009; Lawrence et al., 2013; Vogelstein et al., 2013). This variation can be due to genetic differences, but can also include environmental factors, as well as the cell of origin. For example, endocrine and exocrine
pancreatic cancers exhibit different clinical behaviors (Jiao et al., 2011). Similarly, lung adenocarcinoma and lung squamous cell carcinoma have very different disease courses as well as genomic characteristics (Cancer Genome Atlas Research Network, 2012, 2014; Anusewicz, Orzechowska and Bednarek, 2020).

More recently, heterogeneity within a tumor of a given individual has also been shown. In this case, different cell populations arise and are spatially distinct or intermixed within the tumor and can vary throughout the disease course (Anderson et al., 2011; Landau et al., 2013).

In this section, I will highlight genomic variation, as well as other key drivers of diversity such as transcriptomic and epigenetic changes. Finally, I will describe clinical implications associated with tumor heterogeneity.

1.1.1 Genomic variation

Genomic variation arises throughout the life history of a given tumor through both endogenous and exogenous processes that can lead to single nucleotide variants (SNVs) or chromosomal instability (CIN) (McGranahan et al., 2012; Alexandrov et al., 2015). Some such events can be classified as drivers or founding events, while most will acquire as so-called passenger events (Martincorena et al., 2017; Turajlic et al., 2019).

There are various mutations or copy number aberrations associated with early tumor development. These can be used, for example in breast cancer to classify tumors into subgroups based on hormone receptor expression as well as presence of gene amplifications (Weigelt and Reis-Filho, 2009). Similarly, in lung adenocarcinoma (LUAD) a substantial subgroup of tumors exhibit EGFR-activating mutations, which can be treated with targeted therapies (Pao and Chmielecki, 2010). A separate study also found pervasive gains of chromosome 7p, which encompasses EGFR, in LUAD highlighting that this is a recurrent alteration in LUAD tumors (Watkins et al., 2020). Furthermore, loss of heterozygosity (LOH) of 17p was observed in more than 90% of tumors for HER2+ breast cancer and lung squamous cell carcinoma (LUSC) (Watkins et al., 2020). Chromosome 3p LOH was found in 100% of LUSC tumors investigated in this study (Watkins et al., 2020). Similarly, several genes are recurrently mutated in most pancreatic ductal adenocarcinomas (PDAC), including KRAS, TP53, CDKN2A and SMAD4 (Cancer Genome Atlas Research Network, 2017).
However, although several genes are frequently mutated in various cancer types, ongoing heterogeneity can be observed. Therefore, more recently, studies have aimed to describe intra-tumor heterogeneity (ITH) by sequencing multiple regions of a tumor (Gerlinger et al., 2012; de Bruin et al., 2014; Zhang et al., 2014; Jamal-Hanjani et al., 2017).

1.1.2 Other key drivers of diversity

So far, only genomic drivers of diversity have been mentioned. However, epigenetic and transcriptomic alterations, as well as the tumor microenvironment can also drive heterogeneity within a tumor (Black and McGranahan, 2021).

---

The epigenome plays an important role in regulating cellular processes. Therefore, epigenetic dysregulation, such as DNA methylation, chromatic accessibility and enhancer modifications, can have widespread influences on gene expression (Flavahan, Gaskell and Bernstein, 2017). Such alterations to the epigenome can be transient or fixed; some of these changes are driven by mutations or copy number changes. In several cancer types, mutations in various epigenetic modifiers were found to be driving tumor development (Gröbner et al., 2018).
In addition to variation of gene expression through epigenetic dysregulation, other forms of transcriptomic diversity have been found. These include differential splicing, alternative promoter usage, as well as gene fusions (Kahles et al., 2018; Shiraishi et al., 2018; Demircioğlu et al., 2019; Zhang et al., 2020). Similarly, RNA editing increases transcriptomic diversity within tumors (Baysal et al., 2017).

Finally, the tumor microenvironment presents another key driver of diversity within tumors. On one hand, the immune component driving antitumor immune response plays a key role; on the other hand the non-immune component including tumor-stroma interactions is also a contributing factor (Huebner, Dietzen and McGranahan, 2021). Due to variable levels of immune cells across the tumor, the antitumor immune response and therefore the selective pressures acting on tumor cells will vary. Due to such negative selective pressures acting from the immune system, many cancer cells will lose parts of their antigen-presentation machinery to evade detection and elimination. Such mechanisms include mutations in B2M (Rooney et al., 2015), LOH of the human leukocyte antigen (HLA) (McGranahan et al., 2017), as well as loss or repression of strong neoantigens leading to T cell response (Anagnostou et al., 2017; Rosenthal et al., 2019).

1.1.3 Implications of heterogeneity

Understanding heterogeneity across and within tumors has important clinical implications, such as response to therapy, recurrence of disease, and overall prognosis.

Treatment resistance unfortunately occurs very frequently (J. Gao et al., 2016; Sahin et al., 2017), either through an existing minor subclone which is selected during therapy, or newly acquired alterations leading to resistance to therapies (Mok et al., 2017; Bell et al., 2019; Russo et al., 2019). Understanding the underlying mechanisms better might help inform subsequent therapies to target the emerging variants leading to initial treatment resistance. One such approach includes evolutionary-aware strategies. As treatment resistance does not confer a fitness advantage in the absence of treatment (Siravegna et al., 2015; Xue et al., 2017; Gopal et al., 2019), treating cancers using a chronic approach by temporarily withholding treatment would lead to sensitive cells dividing more frequently than their resistant counterparts. Modeling has shown this approach to be effective (Acar et al., 2020), however to date this has not yet been reproduced in clinical trials (Algazi et al., 2020).
By understanding the extent of ITH, one might be able to gather further insight into the metastatic process. This is explored in detail below in Section 1.3. Finally, intratumor heterogeneity through both treatment resistance and recurrence can have important prognostic implications, with several studies showing relationships between heterogeneity and poor outcome (Landau et al., 2013; Andor et al., 2016; Rye et al., 2018; Turajlic, Xu, Litchfield, Rowan, Horswell, et al., 2018; Iacobuzio-Donahue, Litchfield and Swanton, 2020).

1.2 Tumor development as an evolutionary process

In recent years it has been understood that tumor development is an evolutionary process which follows the principles of evolution by natural selection. This concept was first proposed by Nowell in 1976 (Nowell, 1976), but has since been supported widely in the field (Greaves and Maley, 2012; Yates and Campbell, 2012; Basanta and Anderson, 2013; Vogelstein et al., 2013; Greaves, 2015; Davis, Gao and Navin, 2017). The three main principles of evolution are variation, heredity, and selection. Due to tumor heterogeneity, which is covered extensively in Section 1.1, genotypic and phenotypic variation can be observed within each tumor, with cells differing from one another. As described in Section 1.1.1, genetic alterations account for a large proportion of variation in a tumor. These genetic alterations include single nucleotide variants (SNVs) and somatic copy number aberrations (SCNAs) which are then passed to subsequent generations of cells making these traits heritable. While other key drivers of diversity, such as transcriptomic diversity and epigenetic regulation are not necessarily heritable.
In the following section I will focus on the final principle of evolution by natural selection: selection, or the fitness of a certain phenotype which can be measured through rates of survival or reproduction. Furthermore, I will outline various existing methods to measure selection. Finally, I will explore phylogenetic reconstruction and highlight some existing methods and limitations.

1.2.1 Selection

Selection is one of the main principles of evolution by natural selection and describes the process by which the survival of a specific phenotype is determined. When considering tumor evolution, the unit of selection is at the level of the cancer cell and the evolutionary time period consists of the lifetime of the tumor. During this time period, both positive and negative selection can occur. Positive selection in cancer can arise through
the accumulation of beneficial genetic alterations such as somatic mutations or copy number aberrations that can lead to a fitness advantage relative to other cells within the tumor and can therefore lead to the expansion of certain subclones (Martincorena et al., 2017). A full clonal sweep can occur if one such subclone is significantly fitter than its neighboring cells and therefore has a higher rate of reproduction ultimately leading to the subclone outcompeting neighboring cells. Such a clonal sweep leads to a homogeneous cell population dominating the cancer and therefore a reduction of diversity.

Conversely, negative, or purifying, selection confers a fitness disadvantage which can ultimately lead to cell death. This can occur through accumulation of fitness-reducing genetic alterations in the context of a given environment. For example, the accumulation of potent neo-antigens which can be targeted and removed by the immune system (Marty Pyke et al., 2018; Zapata et al., 2018). Measuring negative selection is possible through identification of an absence of deleterious mutations (Greenman et al., 2006; Martincorena et al., 2017). In order to confidently detect lower mutation density, large cohorts and comprehensive controls for biases, such as germline single nucleotide polymorphism (SNP) contamination are necessary, therefore determination of negative selection is difficult (Martincorena et al., 2017). Furthermore, if both positive and negative selection occur across different mutations in the same population, no clear evidence of selection might be observed. This could lead to underestimation of signals of both positive and negative selection (Martincorena et al., 2017).

1.2.2 Neutral evolution

Neutral evolution can be seen as evolution in the absence of selection. If cells evolve neutrally, they all survive and reproduce at similar rates meaning that the subclones will grow in a proportionate manner and the mutations acquired will only reflect the underlying growth rate rather than any additional fitness advantage (Griffiths, 2003). However, in the life history of a tumor both neutral evolution and selection can play a role at various timepoints. For example, after a clonal sweep occurring due to positive selection, the cells might then evolve predominantly neutrally, with some cells subject to negative selection. In the context of neutral evolution, genetic drift plays a significant role, especially when the population size is small. Genetic drift characterizes the naturally occurring fluctuations of variation between subclones due to random birth-death processes.
1.2.3 Signatures of selection

The accumulation of genetic alterations leading to selection can occur gradually, such as mutations acquired at a steady mutation rate during each cell division (Alexandrov et al., 2015, 2020). Conversely, the accumulation of genetic alterations can also occur in bursts, for example mutations that occur due to mismatch repair deficiency or APOBEC where during a single cell division many mutations might be acquired at once (Hellday, Eshtad and Nik-Zainal, 2014; Nik-Zainal and Morganella, 2017; Petljak et al., 2019). This non-gradual accumulation of alterations can also be referred to as punctuated evolution and can result in rapid expansion of clones and is likely induced by large-scale genomic alterations such as whole genome doubling events or translocations (Baca et al., 2013).

Due to selective pressures, populations might evolve in a way that they converge upon the same phenotype through different mechanisms. This process can be referred to as parallel, or convergent, evolution. Parallel evolution can occur through different mutations arising on separate branches, or through a combination of various genomic or transcriptomic alterations. For example, this could manifest through LOH on one branch and a deactivating mutation on the other, or two separate LOH events targeting alternate alleles (Gerlinger et al., 2012; Jamal-Hanjani et al., 2017; Watkins et al., 2020).

Additionally, to overcome selective pressures, multiple hits of the same gene within the same cell population might occur. This is particularly pronounced in tumor suppressor genes where the tumor cells deactivate both copies of the gene by either nonsense mutations or LOH. Another scenario of such an event would be a pathogenic germline variant in for example one of the BRCA genes and a subsequent mutation or LOH event targeting the other allele.

1.2.4 Different tumor growth models

Tumor evolution is always branched even in the absence of selection, as variation constantly occurs during the life history of each tumor which leads to multiple cell populations being present (Greaves, 2015; Davis, Gao and Navin, 2017; Hu, Sun and Curtis, 2017). Two main models have been proposed which could lead to tumor development and growth, the big bang model and the selective sweep model (Sottoriva et al., 2015; Hu, Sun and Curtis, 2017). The big bang model assumes that initially, punctuated evolution occurs which leads to tumor transformation. However, after tumor
initiation the cell populations grow under neutral evolution leading to intermixed subclones (Sottoriva et al., 2015). In the selective sweep model, after tumor initiation, continuous selection can occur resulting in clonal expansions. Multiple subclones can be present at any given timepoint and compete during tumor growth (Hu, Sun and Curtis, 2017).

1.2.5 Methods to measure selection

Several models which aim to quantify selection and model neutral evolution have been proposed.

![Figure 1-3 Cartoon visualizing the dN/dS method.](image)

*Genes under positive selection have a greater ratio of nonsynonymous (missense and truncating mutations) to synonymous mutations. Genes under negative selection exhibit fewer missense or truncating mutations than would be expected by chance. This figure is adapted from (Martincorena et al., 2017; Black and McGranahan, 2021).*

One method to quantify selection has a long history in the field of species evolution (Nei and Gojobori, 1986; Goldman and Yang, 1994; Yang and Bielawski, 2000). The dN/dS method assumes that if no selection is occurring, the rate of nonsynonymous and synonymous mutations should be equivalent. However, as the total number of sites where synonymous and nonsynonymous can occur differs across the genome, the counts need to be normalized for the number of possible sites these mutations can occur at. Factoring in this normalization, the number of nonsynonymous mutations per nonsynonymous site (dN) is proportional to the number of synonymous mutations per
synonymous site (dS), or the ratio (dN/dS) is equivalent to one. If there are more nonsynonymous mutations occurring than expected, i.e., the ratio is greater than 1, then this is indicative of positive selection. Conversely, if there are fewer nonsynonymous mutations than expected, i.e., the ratio is below 1, then negative selection has potentially occurred in the population for that given set of genes.

However, modifications were needed to apply this model to somatic evolution of cancer (Martincorena et al., 2017). The main adaptation was to include a context-specific refinement to account for mutational processes (Yang, Ro and Rannala, 2003; Greenman et al., 2006; Alexandrov et al., 2013), allowing for mutation rate variation based on the context of the mutations.

Quantifying selection using the dN/dS method can be done at the cohort level – using a population of tumors and focusing on specific genes. The method considers SNVs and calculates selection coefficients for a specified gene list, but also for single genes (Martincorena et al., 2017). In addition to only considering point mutations, the main limitation of this method is that it requires and adequate number of mutations for sufficient power in detecting selection. Consequently, larger cohorts of samples, samples with greater number of mutations, or a larger number of genes need to be considered to obtain significant results.

Some methods also exist to quantify selection at an individual tumor level, for example using the number of mutations and their allelic frequencies. The model proposed by Williams et al. focuses on subclonal mutations as they describe the dynamics of tumor growth and evolution (Williams et al., 2016). Assuming a fixed mutation and proliferation rate, the expected number of new mutations can be calculated using the ploidy, the mutation rate, and the proliferation rate, as well as the number of cancer cells. If the number of cancer cells grows exponentially, the total number of subclonal mutations in a fixed time interval can be calculated. It is known that at any point in time, the allelic frequency of each mutation must be the inverse of the number of alleles in the population. If no selection is occurring this frequency will stay constant during further evolution of the tumor and therefore tumor age and mutational frequency are interchangeable in this scenario. For this reason, substituting the number of alleles in the population with the inverse allelic frequency leads to the conclusion that the cumulative number of mutations per allelic frequency is proportional to the inverse allelic frequency. The authors therefore
postulate that if the cumulative number of mutations is not linearly correlated with the inverse allelic frequency ($R^2 < 0.98$), the tumor is undergoing selection. They also apply this method to identify neutrality in cancer evolution finding neutral evolution in about 30% of cancers. However, this interpretation is inconsistent as they use neutrality as the null hypothesis, and failure to reject this null model does not imply the model to be true (Tarabichi et al., 2018). Furthermore, this method was only developed on single region data, so it is unclear what the best approach to adapting the method to multi-region data would be and initial application to each region shows vast heterogeneity between regions of the same tumor. Finally, the authors restrict their method to mutations with a variant allele frequency (VAF) between 0.12 and 0.24 to ensure they only regard subclonal mutations. However, for example due to genome duplication or low purity of the samples, this approach of using a defined interval to select subclonal mutations can lead to misclassification of these mutations.

A further model that was proposed by Williams et al. suggests that the pattern of ITH represented through the distribution of variant allele frequencies can be used to infer tumor evolution (Williams et al., 2018). They use a branching process model where cells divide and die according to defined birth and death rates and a set number of new mutations are acquired in each generation. The fitness advantage, or selection coefficient, is the ratio of the growth rate of the fitter clone and the growth rate of the background population. If there is no fitness advantage the distribution of allelic frequencies is characterized by a power-law distribution as the authors described in their previous paper (Williams et al., 2016). If there is selection occurring subclonal clusters are apparent in the distribution of allelic frequencies in addition to the neutral tail. The relative size of the clone can be measured using the mean allelic frequency while the total number of mutations is an indication of the relative age of the clone. Furthermore, the time at which the subclone appeared can be estimated using the number of hitchhiking mutations, the number of cell divisions and the mutation rate, which can be estimated using the neutral tail of the distribution. Finally, Bayesian model selection can be used to find the model that best supports the data: neutral evolution with no subclones, or non-neutral evolution with one or more subclones.

A final method describing tumor evolution and selection on a patient level was developed by Sun et al (Sun et al., 2017). The authors propose five different measures of ITH created for multi-region data that consider the allelic frequencies of subclonal mutations.
In the paper the authors use an adapted VAF which also takes into account tumor purity and ploidy, but is different to the cancer cell fraction (CCF), as it takes into account the timing of the mutations in relation to copy number aberrations as well as the cellular prevalence of these copy number aberrations within the tumor. The first of the metrics is the proportion of high frequency mutations to all mutations averaged over all regions. The second metric calculates the average over all pairwise comparisons between regions, calculating the proportion of the number of high frequency region-specific mutations to all region-specific mutations. One main hypothesis underlying these two measures is that selection increases the number of high frequency mutations in each region, i.e., clonal illusion is more prevalent in these samples. The third metric is a fixation index measuring the genetic divergence between regions using the VAF and depth of each mutation in each region. The fourth measure is the maximum Kolmogorov-Smirnov distance between the cumulative VAF distribution of two regions averaged over all pairwise comparisons. The final metric is the ratio of the area under the pooled cumulative VAF distribution to the area under the theoretical cumulative distribution when assuming neutral exponential growth.

The authors simulate various modes of tumor evolution: two neutral models and various values for the selection coefficient. They calculated the five ITH metrics for the simulations as well as some multi-region data and using a principal component analysis and support vector machine to learn the decision boundary were able to classify data into either neutrally evolving tumors or tumors undergoing selection. Although the first two metrics produced a clear separation between the quasi-neutral models and those with higher levels of selection, the values for the last three metrics did not differ greatly between the two main modes of evolution. Furthermore, reproducing the classification of tumors as neutral or non-neutral is difficult, as the decision boundary learned on the principal component analysis may change when using a different set of simulated tumors. For this reason, just implementing the first or second metric proposed should give more robust and reproducible results.

1.2.6 Region based phylogenetic reconstruction

To understand the temporal order of somatic mutation acquisition, the phylogenetic relationships between the different cell populations of the primary tumor need to be inferred and understood. Methods to reconstruct phylogenetic trees can also aid in the
metastatic setting to identify the seeding clone shared between the primary tumor and metastasis which in turn can give further insight into the metastatic process.

Phylogenetic reconstruction of species, or individuals of various species, has been done commonly in evolutionary biology (Rannala and Yang, 2003, 2017; Gil et al., 2013; Lemoine et al., 2018; Kozlov et al., 2019). Therefore, many early studies investigating tumor evolution applied such methods to reconstruct phylogenetic relationships (Kim et al., 2015; Gibson et al., 2016; Zhao et al., 2016; Brown et al., 2017; Naxerova et al., 2017; Zhai et al., 2017; Reiter et al., 2020; Tang et al., 2021). Methods to infer the best phylogenetic tree given the data include maximum parsimony and neighbor joining (Anisimova, 2019). Neighbor joining is a distance-based approach which tries to minimize the distance between two neighboring branches (Saitou and Nei, 1987; Anisimova, 2019). Maximum parsimony, on the other hand, scores the number of steps or changes required (Farris, 1970; Fitch, 1971; Anisimova, 2019). Finding the most parsimonious solution first means enumerating a phylogenetic tree space and subsequently scoring each tree and selecting the optimal solution that minimizes the criteria. However, in some cases the most parsimonious tree will underestimate the amount of evolutionary changes occurring between the species (Felsenstein, 1978; Steel and Penny, 2000; Scotland and Steel, 2015).

These existing phylogenetic reconstruction methods are applied to infer the relationships between individuals of various species. In cancer, it has been extensively shown that there is significant intra-tumor heterogeneity, therefore the likelihood of sampling a homogeneous region of the tumor consisting of only a single cell population is very low. Instead, most tumor samples will include multiple cell populations. For this reason, unless performing single cell sequencing, one cannot directly apply state-of-the-art phylogenetic reconstruction methods. Otherwise, one might incorrectly infer evolutionary patterns such as convergent evolution and the introduction of homoplasies where in fact these might only be due to the presence of cell populations across multiple samples.

To properly infer the evolutionary history of a tumor, one must therefore first deconvolve the cell populations present in each tumor sample before attempting phylogenetic reconstruction. One main limitation of methods such as PyClone (Roth et al., 2014) is that they do not cluster cells by their mutational composition, but rather clusters mutations occurring at similar mutational frequencies, which represents a proxy for
mutations that likely occur in the same set of cells. Therefore, if multiple cell populations exist at similar frequencies in the sample, the mutation occurring in these might be incorrectly grouped together. One way to overcome this is through lineage tracing by way of mitochondrial DNA mutations, somatic epigenetic mutations, or whole genome sequencing. The principle of lineage tracing relies on the assumption that the genomes of cells that are closely related will be more similar than the genomes of cells that are more distantly related and this can be used to reconstruct the clonal relationship of these cells (Gabbutt et al., 2022). The rate at which mutations occur will influence the resolution at which biological phenomena can be studied. If mutations occur infrequently then it is not possible to detect recent clonal divergence, while if mutations occur at high rates it is possible that different lineages will independently develop the same mutations over time.

Such methods have been applied to uncover clonal dynamics in liver, stomach, breast, oesophagus, prostate, and colon (Greaves et al., 2006; McDonald et al., 2008; Fellous et al., 2009; Gaisa et al., 2011; Nicholson et al., 2012; Cereser et al., 2018; Gabbutt et al., 2022). As nuclear DNA has a low somatic mutation rate, using these mutations for lineage tracing is generally infeasible. Mitochondrial DNA mutations occur at rates which are orders of magnitude higher than mutation rates of nuclear DNA. Similarly, DNA methylation changes occur more frequently than mutations in the nuclear DNA. Therefore, lineage tracing is not possible using bulk whole-exome sequencing data from spatially mixed populations and other approaches need to be considered.

1.2.7 Clone based phylogenetic reconstruction

Mutations are genomic scars that accumulate throughout a person’s lifetime and are passed down from cell to cell during division. The infinite sites assumption states that given (nearly) infinite sites exist, the probability of the same site being altered twice is (nearly) zero (Kimura, 1969). Therefore, under the infinite sites assumption, cells sampled from a tumor at present day will contain a complete history of all mutations that have occurred in the lifetime of that tumor. Mutations present in all cells that were sampled likely occurred early, while mutations present only in a subset of cells must have occurred later in the evolution of the tumor. These assumptions can be used to group mutations that occur in a similar proportion of tumor cells meaning they likely occurred at a similar point in the evolutionary history of the tumor (Miller et al., 2014; Roth et al.,
Subsequently, these mutation clusters can be used to infer the cell populations that were present throughout the evolution of the tumor.

However, standard phylogenetic reconstruction methods are not easy to apply even on the deconvolved data as by nature of somatic evolution, mutations leave footprints along the genome. Therefore, not only present-day cell populations are observed, but rather all parental cell populations can be reconstructed. To fully understand which cell populations are present in the tumor samples, one must first deconvolve all cell populations, reconstruct the tree and based on the cellular frequency of mutations as well as their phylogenetic relationships one can then infer which populations are truly present at point of sampling. However, this is highly dependent on the tree topology selected. For this reason, even applying phylogenetic reconstruction methods that allow for fossil populations to be included which can aid in timing divergence of populations is not possible as one needs to classify which are present-day and which are fossilized populations (Benton and Donoghue, 2007; Heath, Huelsenbeck and Stadler, 2014; Álvarez-Carretero et al., 2022).

One limitation in phylogenetic reconstruction of cancer cell populations are variable mutation rates occurring due to different mutational processes acting during tumor evolution which could lead to long branch artefacts. In such instances the higher number of mutations might not be indicative of evolution over longer periods of time, but rather mutations occurring in bursts. Therefore, timing events during tumor evolution using mutation rates could lead to incorrect inference. However, by clustering mutations occurring at similar frequencies using methods such as PyClone (Roth et al., 2014) and subsequently inferring their evolutionary ordering by reconstructing a directed graph of such clusters, variable mutation rates should not bias such approaches.

1.2.8 Sampling bias during clonal reconstruction

Cancer evolution by definition is a branched process, as through selection or genetic drift variation will be present in the cell population at any given point in time. However, clonal sweeps can occur, leading to a reduction in heterogeneity within the tumor and therefore incorrect inference of linear evolution (Davis, Gao and Navin, 2017). Sampling bias might overestimate this observation further.
When only sampling a single region of a primary tumor it is possible that a clonal sweep has occurred and therefore parental cell populations cannot be inferred correctly. Although all mutations that have occurred throughout the life history of the lineage can be observed, there is not enough resolution to properly infer the different cell populations and therefore all mutations appear clonal (clonal illusion) (de Bruin et al., 2014). Similarly, if the sample contains multiple but not all cell populations present in the primary tumor, this could lead to incorrect inference of linear evolution.

By sampling multiple regions this sampling bias can be mitigated. However, it is still possible that due to the limit of detection some mutations and therefore minor subclones are missed. Therefore, by integrating metastasis samples and leveraging mutations at lower frequency, a more complete picture of tumor evolution through space and time can be obtained.

1.3 Evolutionary trajectories in metastasis

In order to characterize metastatic evolution, one must first differentiate between divergence and dissemination of metastatic competent cells. Divergence of metastatic cells from the primary tumor can occur much earlier than the dissemination process that ultimately leads to cells migrating to a metastatic site. Metastatic divergence can be defined as the moment in the evolutionary history of a tumor when the cell population that is subsequently found in metastases differentiates from the remaining cell populations of the primary tumor. Dissemination of this cell population can occur at any point after divergence. Timing metastatic dissemination cannot be done with tissue data alone, but rather might be possible using imaging or circulating tumor DNA (ctDNA).

In this section I will focus on definitions and methods to time divergence of metastases as well as characterizing the dissemination patterns. Furthermore, I will highlight multiple paired and unpaired metastasis studies that have investigated the metastatic process.

1.3.1 Timing metastatic divergence

Metastases have been shown to diverge and disseminate from the primary tumor both early and late (Turajlic and Swanton, 2016). Understanding and classifying the timing of metastatic spread has important clinical implications. If metastases generally occur early, early detection likely will not significantly improve patient outcomes. Furthermore, in the
In the case of early divergence of metastases it is possible that not all primary mutations are shared with metastases meaning that therapies targeting specific variants in the primary tumor might not be effective in treating metastases. Conversely, if metastases occur late in tumor evolution, the tumor could be detected and treated before metastases occurred and most primary mutations would also be shared with the metastases (Turajlic and Swanton, 2016; Lee et al., 2020).

Figure 1-4 Schematic showing timing of divergence.
This figure is adapted from our Cell SnapShot (Huebner, Dietzen and McGranahan, 2021) and shows a representative example of when metastases might diverge and what their cell composition could look like. In the case of early divergence, the cells shared between the primary tumor and metastasis occurred early in tumor evolution, and subsequently a complete clonal sweep occurred within the primary tumor that is not shared with the metastases. In the case of late divergence, the cells shared between the primary and metastasis have arisen later in tumor evolution and there has been no additional clonal sweep in the primary tumor.

Timing metastatic divergence and dissemination in absolute time is difficult. One option to time metastases is by imaging, however in most cases the metastasis needs to reach
sufficient size before it is detectable using imaging techniques. Therefore, this method might underestimate the length of time the metastasis existed.

A second approach to timing metastases is using ctDNA (Tie et al., 2016; Wan et al., 2017; Abbosh, Birkbak and Swanton, 2018). If tumor DNA can be found in blood this can be an indication of cells disseminating from the primary tumor (Abbosh et al., 2017; Phallen et al., 2017; Cohen et al., 2018; Christensen et al., 2019; Reinert et al., 2019). However, not all tumors shed ctDNA in the blood and so the sensitivity is lacking (Abbosh et al., 2017). Furthermore, not all dissemination will occur through the bloodstream and other methods of cell migration have been reported, such as spread through airspace (STAS) (Toyokawa et al., 2018; Yokoyama et al., 2018; Mino-Kenudson, 2020) or lymphatic spread.

Using genomic data, one can time metastases using mutations. Assuming mutations accumulate at constant rate in the life history of a tumor, it is possible to quantify the number of additional mutations the metastasis has acquired compared to the primary tumor. As not all mutations accumulate at constant rate (Nik-Zainal and Morganella, 2017; Petljak et al., 2019; Alexandrov et al., 2020) and especially chemotherapy agents have been associated with higher mutation burden (Johnson et al., 2014; Murugaesu et al., 2015; Pich et al., 2019, 2021; Landau et al., 2020) only certain mutations in specific contexts can be used to time metastases. However, using whole exome sequencing, not enough such mutations are detected and therefore in most cases this analysis lacks power and requires whole genome sequencing in order to properly estimate the timing. For this reason, using whole exome sequencing data only relative timing of metastatic divergence is possible.

Timing relative metastatic divergence can be done by comparing similarities of mutations. For example, if the majority of mutations are shared between the primary and metastasis this can be defined as late divergence, while if only a small proportion of mutations are shared this can be defined as early divergence. Using a phylogenetic tree, it is possible to go one step further and use the trunk of the tree to define divergence. If not all truncal mutations of the primary tumor are shared with the metastases, then this can be defined as early divergence, while if the metastasis shares all truncal mutations and diverges after branching of the primary tumor phylogeny has occurred, this can be defined as late divergence (Turajlic and Swanton, 2016). An additional way to relatively
time metastatic divergence is by quantifying the size of the primary tumor at the point of divergence (Hu et al., 2019). If the primary tumor was small, for example below the limit of detection, when the metastases diverged this can be defined as early divergence, otherwise divergence can be classified as late (Hu et al., 2019).

Several studies have found predominantly late divergence. In esophageal cancer metastases were generally found to diverge late, with a few tumors exhibiting early divergence (Noorani et al., 2020). Similarly, in pancreatic cancer predominantly late divergence was also observed with modeling suggesting that metastatic potential was acquired 5 years after the original parental cell existed (Yachida et al., 2010). In breast cancer, late divergence was observed with metastatic divergence occurring on average at 87% of the molecular age of the primary tumors (Yates et al., 2017).

Conversely, multiple studies, including breast and colorectal cancer, have found predominantly early divergence of metastases (Faltas et al., 2016; Brown et al., 2017; Hu et al., 2019). However, these findings might be confounded by the use of region-based trees used to infer the timing, as well as under-sampling of the primary tumor.

### 1.3.2 Modes of metastatic dissemination

A second important characterization of metastases is their mode of dissemination (Birkbak and McGranahan, 2020). If a single cell population is shared between the primary tumor and metastases (i.e., monoclonal dissemination), this could mean that metastatic potential was acquired once during tumor evolution and therefore by understanding drivers of this metastatic potential treatments targeting these drivers could be applied. If multiple cell populations are shared between the primary tumor and sites of metastases (i.e., polyclonal dissemination), this could mean that metastatic potential was acquired early on in tumor evolution, so called “born to be bad” tumors. A second possibility in this case is that metastatic potential was acquired multiple times in parallel which could mean combination therapies targeting multiple events are necessary to treat such tumors.

However, one important limitation when defining modes of metastatic dissemination, especially using tissue samples, is that only a snapshot of the tumor evolution is captured. This means that for example, multiple cell populations could have migrated to a metastatic site, but if one cell population had a fitness advantage the observed pattern
would be monoclonal dissemination. Finally, dissemination can not only occur from the primary tumor to a site of metastasis, but rather also one metastasis can seed a subsequent metastasis (El-Kebir, Satas and Raphael, 2018).

Figure 1-5 Schematic showing metastatic divergence. This figure is adapted from our Cell SnapShot (Huebner, Dietzen and McGranahan, 2021) and shows a representative example of metastatic dissemination and what the cell composition of metastases could look like. In the case of monoclonal dissemination, only a single cell population is shared between the primary tumor and metastasis. Conversely, in the case of polyclonal dissemination multiple cell populations are shared between the primary tumor and metastasis. In this case, the pattern can be differentiated further into monophyletic and polyphyletic. In the former case, the multiple shared cell populations occur along one branch of the phylogenetic tree and are therefore all children or ancestors of one another. While in the case of polyphyletic dissemination, the multiple cell populations arise as siblings or cousins of one another.

In order to infer monoclonal or polyclonal dissemination of metastases, clonal deconvolution needs to be determined and clone-based phylogenies reconstructed. Therefore, many studies that have explored dissemination using region-based phylogenetic reconstruction and consequently underestimated the heterogeneity of individual samples are not comparable to the approach taken here (Kim et al., 2015;
Gibson et al., 2016; Zhao et al., 2016; Brown et al., 2017; Naxerova et al., 2017; Zhai et al., 2017; Reiter et al., 2020; Tang et al., 2021). Nonetheless, predominantly monoclonal dissemination has been observed in multiple studies. For example, brain metastases derived from multiple primary tumor origins were found to be monoclonal (Brastianos et al., 2015). Similarly, monoclonal dissemination was observed in lung cancer and clear cell renal carcinoma among other cancer types (Turajlic, Xu, Litchfield, Rowan, Chambers, et al., 2018; Hu et al., 2020; Lee et al., 2020). In ovarian cancer, as well as breast cancer and colorectal cancer, both monoclonal and polyclonal seeding was observed (Hoadley et al., 2016; McPherson et al., 2016; Brown et al., 2017; Angelova et al., 2018; Ullah et al., 2018; De Mattos-Arruda et al., 2019).

1.3.3 Metastasis-to-metastasis seeding

The presence of lymph node (LN) disease at diagnosis is associated with higher stage and poorer outcome in multiple cancer types (Goldstraw et al., 2016). One explanation might be that lymph nodes are a gateway to subsequent local and distant metastases, therefore serving as an intermediate step to disease recurrence. However, it is also possible that lymph node metastases simply represent the metastatic potential of a tumor.

Several studies of different tumor types have investigated the role of lymph node metastasis and distant metastases. In colorectal cancer, the distant metastases were found to be seeded directly from the primary tumor in the majority of cases, however in a subset seeding from the lymph node was observed (Naxerova et al., 2017). In a study focusing on breast cancer, the lymph node metastases did not seem to be involved in the seeding of distant metastases (Ullah et al., 2018). Similar results were shown in a extensively sampled prostate cancer patient where the metastases were not derived from the lymph node metastases (Haffner et al., 2013) as well as in esophageal cancer, where the distant metastases shared more similarities to the primary tumor than lymph node metastases (Noorani et al., 2020). In a recent study of lung cancer metastases, similar patterns were observed (Tang et al., 2021). Finally, several studies of different primary tumor origins have found evidence of metastasis-to-metastasis seeding (Gundem et al., 2015; Brown et al., 2017; El-Kebir, Satas and Raphael, 2018; Chen et al., 2022).
1.3.4 Metastasis driver events

As genomic changes are heritable, each metastatic cell will also include changes that accumulated during early tumor development and might not be directly linked to the metastatic process. Therefore, to investigate events driving formation and dissemination of metastases (metastasis driver events), ideally paired primary-metastatic cohorts are needed. In these cases, primary unique, metastasis unique and shared genomic events can be classified.

However, the majority of these studies have only included small cohorts and therefore lack power to fully investigate metastasis driver events. Instead, large-scale cohort analyses exist that have compared primary and metastatic samples in unpaired analyses (Robinson et al., 2017; Bertucci et al., 2019; Priestley et al., 2019; Nguyen et al., 2022). However, to date no new metastasis unique drivers have been identified. A recent large-scale study of unpaired metastatic tumors showed that in a large number of tumor types, including lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), the metastases were significantly more chromosomally unstable compared to primary tumors (Nguyen et al., 2022). Similarly, a higher fraction of whole genome doubling (WGD) as well as higher tumor mutational burden (TMB) was observed in metastases in some cancer types, including LUAD. Furthermore, metastases across multiple tumor types were found to be more homogeneous than primary tumors, possibly indicating a bottlenecking event occurring (Nguyen et al., 2022). Finally, the frequency of recurrent oncogenic alterations was investigated, with several alterations found to be more frequent in metastases. These events included TP53 mutations as well as CDKN2A deletions, PTEN mutations and deletions, and MYC amplifications (Nguyen et al., 2022).

However, as mentioned above due to the unpaired nature of this analysis it is not possible to infer whether these events are metastasis-specific events or rather enriched in the primary tumors of patients who develop recurrence of their disease. In the same study, the authors explored differences between metastatic and non-metastatic tumors and found that chromosomal instability as well as TMB was higher in primary tumors that developed recurrence. Furthermore, TP53 mutations were found to be more frequent in primary tumors of metastatic patients (Nguyen et al., 2022). Taken together, these results could indicate that these events are likely enriched in metastasizing tumors and subsequently over-represented in metastatic lesions.
A study analyzing paired and unpaired metastatic samples across LUAD, breast cancer and clear cell renal cell carcinoma found several focal somatic copy number aberrations (SCNAs) enriched in metastatic samples (Watkins et al., 2020). In LUAD, a loss of chromosome 17p encompassing NCOR1, NF1, and TP53 was found to be maintained in paired analysis, with the frequency of this loss event enriched in metastasis in an unpaired analysis. Similarly, a loss of chromosome 19p including DAZAP1 and STK11 was predominantly maintained in the paired analysis and enriched in metastasis in an unpaired analysis (Watkins et al., 2020).

By leveraging matched primary metastatic samples, events can be investigated that are enriched in metastatic sites compared to the primary tumor. Several studies have shown that the majority of driver alterations occur early in tumor evolution, on the trunk of the phylogenetic tree and are therefore shared with metastases (Makohon-Moore et al., 2017; Yates et al., 2017; Reiter et al., 2018; Turajlic, Xu, Litchfield, Rowan, Horswell, et al., 2018). However, other studies have found metastasis-unique drivers that could potentially be clinically targetable (Brastianos et al., 2015).

Overall, the data from both unpaired and paired analyses of metastatic disease support findings that the majority of driver alterations are acquired early during tumor evolution and are common across metastases with only few novel metastasis unique events being observed.

1.4 Tracking tumor evolution over time using ctDNA

Circulating tumor DNA (ctDNA) is extracted from a patient’s blood sample, specifically the plasma and refers to the fraction of cell-free DNA that is derived from the tumor (Abbosh, Birkbak and Swanton, 2018). This represents a minimally invasive approach to measure tumor markers for early diagnosis as well as minimally residual disease (MRD) post initial diagnosis (Heitzer et al., 2013; Tie et al., 2016; Abbosh et al., 2017).

As mentioned in other sections, ctDNA can be used to infer timing of metastatic dissemination as well as modes of dissemination (Burgener et al., 2017). In patients whose tumors shed ctDNA, it presents an almost perfect proxy for metastatic detection. However, not all tumors shed ctDNA. This likely reflects histology specific differences,
specifically in lung adenocarcinomas only about half of the tumors show evidence of tumor DNA in blood (Abbosh et al., 2017).

The main benefit of using ctDNA to track tumor evolution is the fact that it is minimally invasive and therefore this approach can be applied much more frequently than tissue sampling. Furthermore, sampling bias might play a less significant role in ctDNA sampling compared to tissue sampling where only a small portion of a tumor is sampled and additionally, not all metastatic tumor sites can be sampled due to patient frailty or difficulty in accessing the site of metastasis (Parikh et al., 2019). A single patient case study found that the majority of mutations found in ctDNA overlapped with those found in the primary tumor and metastatic tissue samples. The ctDNA samples highlight differences in treatment response over the time period sampled (Murtaza et al., 2015).

Conversely, some main limitations include that ctDNA shedding is linked with tumor volumes (Abbosh et al., 2017). Furthermore, not all tumors shed ctDNA even at high tumor volumes meaning the sensitivity of this approach is limited. Finally, even in cases where ctDNA is present, ctDNA content is often quite low, limiting the confidence in detecting specific variants, and more so detection of new variants is very limited (Morelli et al., 2015; Murtaza et al., 2015; Abbosh, Birkbak and Swanton, 2018; Tjensvoll et al., 2022).

Previous studies have therefore made use of targeted detection of mutations in frequently mutated cancer genes (Lanman et al., 2015; Newman et al., 2016). Other studies have created bespoke patient-specific panels and performed multiplex-PCR next-generation sequencing to track mutations (Abbosh et al., 2017). However, depending on variants selected, specifically subclonal variants or incorrectly inferred truncal variants, mutations might not be picked up in the ctDNA samples.

Adalsteinsson et al. developed a method to quantify the tumor content in cfDNA. The method, ichorCNA, uses ultra-low pass whole genome sequencing to detect presence of ctDNA in a sample without prior knowledge of mutations (Adalsteinsson et al., 2017). The tumor fraction was estimated and if sufficient tumor derived DNA was present in the cell-free DNA, subsequent whole exome sequencing (WES) could be performed to analyze the genetic diversity of such samples. The whole exome sequenced samples were subsequently compared to the matched tumor biopsies to infer overlap of the
cfDNA and tissue samples. The tool predicts copy number aberrations as well as tumor fraction while accounting for ploidy.
Chapter 2  Data and Methods

2.1 Data

2.1.1 TRACERx421 paired primary metastatic cohort

TRACERx (TRAcking non-small cell lung Cancer Evolution through therapy [Rx]; ClinicalTrials.gov number NCT01888601) is a national, prospective clinical study investigating early-stage non-small cell lung cancer (Jamal-Hanjani et al., 2014, 2017) and was approved by an independent ethics committee (13/LO/1546). The study aims to recruit 842 patients with stage I-III (TNM version 7) disease which are then followed over a period of five years with regular blood samples taken. In the unfortunate case of disease recurrence, sampling of the metastatic site is also attempted.

The aim of TRACERx is to investigate the impact of tumor heterogeneity on prognosis, as well as the effect of treatment on intra-tumor heterogeneity. For this, multiple sites of the primary tumor are sampled after surgery with curative intent. Additionally, any lymph nodes extracted during surgery (primary lymph nodes) harboring tumor tissue are also sampled. Dual DNA and RNA extraction is performed on fresh frozen tissue from each sample, and the DNA samples are whole exome sequenced, aiming for a median coverage of 400x per sample.

The TRACERx421 cohort contains the first 421 patients recruited to the study, including the 100 patients first described in (Jamal-Hanjani et al., 2017). Some patients presented with multiple synchronous primaries at surgical resection, therefore the total number of tumors included in the cohort is 432 with a total of 1553 primary tumor regions analyzed.

For the paired primary metastatic cohort, all patients within the TRACERx421 cohort were included, if they either presented with lymph node (LN) disease at primary surgery or developed subsequent recurrence or progression of their disease. In total, 129 patients from the 421 were included in the metastatic cohort, with 227 metastatic regions analyzed. Of these, 99 patients had primary LN disease (123 formalin-fixed paraffin embedded [FFPE] and 32 fresh frozen LN samples) and 50 patients had recurrent/progressive disease (18 FFPE and 51 fresh frozen samples). There was an overlap of 21 patients with both primary LN and recurrent/progressive disease.
Additionally, there were two patients with synchronous lesions detected at surgery (3 fresh frozen samples), one of which also had primary LN disease.

2.1.2 PEACE cohort

PEACE (Posthumous Evaluation of Advanced Cancer Environment; ClinicalTrials.gov number NCT03004755) is a national, prospective clinical study. Metastatic patients with tumors of various tissues of origin are recruited to the study and when they pass away, research autopsies are performed, and the breadth of metastatic sites are sampled. PEACE allows for extensive sampling across multiple metastatic sites to recreate a more complete picture of metastatic disease. A subset of patients recruited into PEACE were previously recruited into TRACERx, allowing multi-timepoint sampling of at least 2-3 timepoints during disease relapse and progression.

The aim of PEACE is to explore metastatic disease and understand why patients die. One main objective is studying cancer evolution in the context of the tumor microenvironment and immune landscape. Secondly, the aim is to take a multi-omics approach in understanding the processes involved in metastatic disease.

To date, 22 autopsies of patients co-recruited in TRACERx and PEACE have been performed and a total of 476 tumor samples sequenced. This includes 110 surgical samples (97 primary tumor samples and 13 primary LN), 13 recurrence samples, 5 progression samples, and 348 samples from tissue harvest. For the primary LN, fresh frozen tissue was sequenced for 4 samples, while FFPE tissue was used for the remaining 9 samples. One of the recurrence samples was sequenced from FFPE tissue, with the remaining recurrence and progression samples all derived from fresh frozen tissue. Tissue harvest samples were all sequenced from fresh frozen tissue. Finally, two patients within the PEACE cohort were initially recruited into TRACERx, but subsequently excluded from further analysis, as they did not meet the inclusion criteria. This was due to the fact that all tumor samples sequenced from fresh frozen tissue did not pass mutation and copy number quality control. For these patients, laser capture macro/micro-dissection of FFPE tissue blocks was performed to extract DNA from the primary tumor (see also Section 2.2.1.1). Therefore, for these two patients (4 primary tumor samples) FFPE tissue was sequenced. The remaining 93 primary tumor samples were all sequenced from fresh frozen tissue.
2.1.3 Pancreatic ductal adenocarcinoma cohort

This cohort consists of a total of 24 patients with pancreatic ductal adenocarcinoma. 20 of these were enrolled in a multicenter, randomized phase II clinical trial (ClinicalTrials.gov number NCT02795650), with an additional 4 patients recruited at the Vall d’Hebron Institute of Oncology. The patients recruited as part of the study underwent a tumor biopsy, including a derived PDX model, before starting first-line treatment and received personalized treatment as second or third-line therapy. The main aim of the study is to understand whether personalized treatment improves one year survival of the patients compared to the control arm receiving standard-of-care treatment.

Overall, 87 tumor samples were sequenced for the 24 patients. This included 2 regions of the primary tumor for one patient and 22 metastatic biopsies. Furthermore, 8 PDX samples were also sequenced. Finally, 55 cfDNA samples were subject to whole exome sequencing.

2.2 Methods

2.2.1 Processing of TRACERx and PEACE cohorts

Data processing of the TRACERx and PEACE cohorts is broadly consistent with the pipeline described in (Jamal-Hanjani et al., 2017) as well as in the manuscript “TRACERx: The evolution of metastases in non-small cell lung cancer” by the TRACERx consortium currently under review in Nature.

2.2.1.1 Sample extraction and whole exome sequencing

For germline DNA extraction, the QiaAmp blood midi kit (Qiagen) was used on peripheral blood (2ml). For tumor samples, dual extraction of DNA and RNA was performed using a modified DNA/RNA AllPrep kit (Qiagen). For samples acquired at surgery this was done on 3x3x3mm of tumor tissue, while for samples obtained from core biopsies or endobronchial ultrasound (EBUS) biopsies most of the tissue was used. For FFPE tumor samples, laser capture macro/micro-dissection was performed on slides obtained from FFPE tissue blocks, and subsequently DNA was extracted using the Qiagen GeneRead FFPE DNA kit.
A modified version of the Agilent Human All Exome V5 kit was used to perform exome capture. Samples were subject to paired-end multiplex sequencing on the Illumina HiSeq 2500 and HiSeq 4000. For all fresh frozen samples sequenced, including recurrence/progression samples, the paired germline sample (extracted before the initial surgery) was re-sequenced to adjust for any sequencing-specific artefacts. For FFPE samples the paired germline was not re-sequenced.

### 2.2.1.2 Alignment

For each lane of sequencing, the paired-end reads were aligned separately using BWA-MEM (version 0.7.17) (Li and Durbin, 2009). Subsequently, all sequencing lanes of a single sample were merged using Sambamba (version 0.7.0) (Tarasov et al., 2015) and deduplicated using Picard Tools (version 2.21.9; http://broadinstitute.github.io/picard/). The Genome Analysis Toolkit (GATK, version 3.8.1) (McKenna et al., 2010) was used for local realignment around INDELs. Finally, quality control was performed using Somalier (version 0.2.7; https://github.com/brentp/somalier), Samtools (version 1.9) (Li et al., 2009), Picard Tools, and Conpair (version 0.2).

### 2.2.1.3 Somatic mutation calling

Variant calling was performed analogously to (Jamal-Hanjani et al., 2017). In the first step, non-reference positions were located in both tumor and germline samples using SAMtools mpileup (v1.10). Bases with a phred score <20 and reads with a mapping quality <20 were removed. Subsequently, VarScan2 somatic (v2.4.4) (Koboldt et al., 2012) was used to characterize somatic variants between the tumor samples and their matched germline. VarScan2 was run using default parameters with the exception of the germline coverage, which was set to 10, the minimum variant frequency, which was set to 0.01, and tumor purity which was fixed as 0.5. Consequently, the somatic mutation calls were filtered using VarScan2's fpfilter.pl script, once using default settings and again setting the min-var-frac to 0.02. In addition to VarScan2, MuTect (v1.1.7) (Cibulskis et al., 2013) was also implemented to infer somatic mutations and variants were filtered based on the “PASS” parameter. Finally, analogously to (Jamal-Hanjani et al., 2017) further filtering was performed to minimize false positive calls.

A variant was defined as a true SNV if the variant allele frequency (VAF) was greater than 5% and the mutation was called by VarScan2 with a somatic p-value ≤ 0.01.
Alternatively, if the VAF was greater than 2%, the mutation was considered a true positive if it was called by both VarScan2 (somatic p-value ≤ 0.01) and MuTect. Furthermore, the coverage at the locus needed to be ≥30 in all tumor regions, and, due to filtering of low-frequency artefactual mutations, the variant count needed to be ≥10. Conversely, in the germline the variant needed to be associated with a VAF ≤ 1% and a variant read count <5. Finally, any variants found to have a population frequency within the TRACERx cohort >1% were also filtered.

Insertion/deletion (INDEL) calls from VarScan2 were filtered in the same way described above, but with a higher coverage of ≥50, and a somatic p-value ≤ 0.001 needed to infer an INDEL as a true positive. If any discrepancies in start, or end positions, or length of the INDEL were detected across multiple regions, the longest INDEL was reported.

Finally, the multi-region sequencing was leveraged to determine presence/absence of variants across regions. Firstly, all detected variants across all regions were extracted and subsequently the variant and reference counts as well as coverage were re-calculated for all variants across all regions using bam-readcount (v0.8.0) (https://github.com/genome/bam-readcount). Any variant detected at the filters described above in at least one region was classified as a high confidence variant and subsequently, the filters to define such a variant in other regions were reduced such that the variant was detected in a region at a VAF ≥ 1% and a variant read count ≥3. If a variant did not pass the original high-confidence filters in any region, it was not included in subsequent mutation analyses.

Variants were annotated using Annovar (Wang, Li and Hakonarson, 2010) and COSMIC (v75) (Sondka et al., 2018).

2.2.1.4 Distinguishing genomically unrelated tumors

To infer whether all samples obtained from a single patient were genomically related and formed a single tumor, ubiquitous mutations were leveraged. A mutation was defined as ubiquitous if it was present at a variant allele frequency (VAF) of at least 1% across all regions. If more than 10 ubiquitous mutations were found, the samples were classified as a single tumor. If, however, less than 10 ubiquitous mutations were detected, the R function hclust was used to cluster the samples based on their mutation VAFs. The clustering tree was cut such that only two branches remained,
and these were classified as separate tumors. The same steps were repeated for each of the two sets of samples, giving rise to up to four distinct tumors.

2.2.1.5 Somatic copy number alteration calling

As previously described in (Jamal-Hanjani et al., 2017), single nucleotide polymorphisms (SNPs) were identified from the germline samples using platypus (v0.8.1) (Rimmer et al., 2014). VarScan2 was used to calculate log-ratio (logR) data, and GC correction was performed using a wave-pattern GC correction method based on (Cheng et al., 2011). Both the SNP and logR data were used as input to ASCAT (v2.3) (Van Loo et al., 2010) and Sequenza (v2.1.2) (Favero et al., 2015) to infer copy number segmentation, as well as tumor purity and ploidy. Subsequently, the data were manually reviewed and samples with low tumor purity (<10%) were excluded from subsequent copy number analysis.

These data were then used as input to a recently described multi-region whole exome sequencing approach (Watkins et al., 2020). The approach leverages allelic imbalance from multi-region data to phase heterozygous SNPs and re-estimate allele-specific copy number across regions, which can help rescue allelic imbalance calls in lower purity samples. Furthermore, due to the phasing of SNPs, mirrored subclonal allelic imbalance (MSAI) can be inferred.

Finally, to account for the increased variation in logR estimates and subsequent segmentation in FFPE samples, additional processing was performed during copy number estimation of FFPE samples. Instead of the logR segmentation used for fresh frozen samples, segmentation based on b-allele frequency (BAF) estimates was implemented and the mean logR value of all SNPs within such a segment was used as the segmented logR value. Furthermore, small segments without any heterozygous SNPs were joined with the adjacent segment with the closest mean logR value. Finally, segments with the lowest logR values (<5% of the sample) were removed.

2.2.1.6 Estimating the mutation copy number and cancer cell fraction

The cancer cell fraction (CCF) and mutation copy number (mutCN) of a mutation are calculated as described in (McGranahan et al., 2015; Jamal-Hanjani et al., 2017). For each mutation, the observed VAF of each mutation is based on the proportion of cancer cells harboring the mutation of interest (i.e., the cancer cell fraction), however...
the relationship is confounded by the normal cell contamination (i.e., tumor purity), as well as the copy number multiplicity. In brief, if a sample consists of 50% tumor cells and 50% normal cells, both with a copy number of 2, a mutation present in all tumor cells will have an overall VAF of 0.25. If the tumor has a higher ploidy of 4 and the mutation only occurs on a single allele of the tumor cells, this VAF will decrease to 0.1667. It is easily apparent that multiple combinations of purity and ploidy can lead to the same VAF estimate. For example, a VAF of 0.25 can occur, as described above, when the tumor purity is 0.5 and the tumor ploidy is 2 and the mutation occurs in all cancer cells on a single allele. Similarly, a VAF of 0.25 can occur if the tumor purity is 1 and the tumor ploidy is 2 and the mutation occurs in half of the tumor cells (CCF = 0.5). Other alternative scenarios are also possible. This leads to the following equation describing the relationship of these variables to estimate the expected VAF, $VAF_e$, of a given mutation:

$$VAF_e = \frac{\text{purity} \times \text{CCF}}{(\text{purity} \times \text{tumorCN} + (1 - \text{purity}) \times \text{normalCN})}$$

where $\text{purity}$ is the fraction of tumor cells in the sample, $\text{tumorCN}$ is the local copy number in the tumor sample, and $\text{normalCN}$ is the local copy number in the normal sample, which is assumed to be 2.

Given tumor purity and ploidy estimates, including copy number estimates of all segments across the genome, the above equation can be used to estimate the cancer cell fraction as well the mutation copy number, which describes the number of alleles the mutation occurs on. Given a mutation with $n_{\text{var}}$ variant reads and a total depth of $N$, the probability of the CCF can be estimated as:

$$P(\text{CCF}) = \text{binom}(n_{\text{var}}|N, VAF_e)$$

The CCF values can be calculated over a uniform grid of 100 CCF values between 0.01 and 1 and normalized to obtain a posterior CCF distribution from which, for example, 95% confidence intervals can also be obtained. Finally, the mutation copy number, $\text{mutCN}$, can be calculated as:

$$\text{mutCN} = \frac{VAF}{\text{purity}} \times (\text{purity} \times \text{tumorCN} + (1 - \text{purity}) \times \text{normalCN})$$

where $VAF$ is the variant allele frequency of the mutation, $\text{purity}$ is the fraction of tumor cells in the sample, $\text{tumorCN}$ is the local copy number in the tumor sample, and $\text{normalCN}$ is the local copy number in the normal sample, which is set to 2.
2.2.1.7 Mutation clustering and tree building

Mutation clustering was performed using a modified version of PyClone (Roth et al., 2014). Details of the modifications implemented are described in Chapter 3.

In short, PyClone is a hierarchical Bayes statistical model which performs Dirichlet Process (DP) clustering of mutations from a single or multiple samples (Roth et al., 2014). Allelic counts of mutations are used as input and prior information is obtained from copy number estimates as well as optional tumor purity estimates. Each mutation is assigned a posterior density of the mutation prevalence and the model also outputs a matrix of probabilities of any two mutations occurring in the same cluster. The model does not cluster cells by their mutational composition, but rather groups mutations occurring at similar mutation frequencies. This approach can be significantly improved when considering multiple samples as well as deep sequencing (Roth et al., 2014).

The method uses a DP prior which converts the continuous Uniform(0,1) distribution to a discrete distribution such that mutations sharing the same mutation prevalence have a non-zero prior probability (Roth et al., 2014). Subsequently, Markov Chain Monte Carlo (MCMC) sampling from the posterior distribution is performed as the exact distribution cannot be inferred. Additionally, five strategies are considered for implementing genotype aware prior distributions. In all five cases, the normal population of cells is fixed as AA (i.e. consisting of two copies of the reference allele). Given total copy number estimate $c_n$ and $c_{n_1}, c_{n_2}$ as copy numbers of the two alleles, the genotypes of the reference ($g_r$) and variant ($g_v$) populations are defined as follows:

1) AB prior: $g_r = AA, g_v = AB$. This means the mutation is assumed to be diploid and heterozygous, while the reference population is diploid with two reference alleles.

2) BB prior: $g_r = AA, g_v = BB$. This means the mutation is assumed to be diploid and homozygous, while the reference population is diploid with two reference alleles.

3) No zygosity prior: $g_r = AA, g_v$ is not fixed, but the copy number of the variant population, $cn(g_v) = cn_1$ with exactly one mutant allele $b(g_v) = 1$.

4) Total copy number prior: $g_r = AA$ or $g_r = A...A$, with $cn(g_R) = cn_1$; $g_v$ is not fixed, but the copy number of the variant population, $cn(g_v) = cn_1$ with $b(g_v)e\{1,...,cn_1\}$. 


5) Parental copy number prior: $g_v$ is not fixed, but the copy number of the variant population, $cn(g_v) = cn_t$ with $b(g_v)\in\{1,cn_1, cn_2\}$. If $b(g_v)\in\{cn_1, cn_2\} g_r = AA$, while if $b(g_v) = 1$ then $g_r = A...A$, with $cn(g_R) = cn_t$. The former assumes the mutation occurred before the copy number event while the latter assumes that the copy number event occurred first. (Roth et al., 2014) showed that the parental copy number prior performed best on simulated data followed by the total copy number prior with the other three priors which do not take into account mutational genotypes performing worse. This highlights the importance of incorporating the genotype of mutations when performing mutation clustering. However, one limitation of this method consists of the assumption that all cells have the same genotype. Additionally, the method cannot correct for mutation loss due to loss of heterozygosity of one allele. Finally, using simulations the authors show that with increasing number of mutations PyClone over clusters the mutations and the overall performance deteriorates (Roth et al., 2014).

For this reason, an adapted version of PyClone was implemented and described in Chapter 3 to overcome the limitations of the method due to the number of mutations by reducing the number of mutations clustered simultaneously; as well as variations of mutation frequency due to subclonal copy number aberrations.

### 2.2.1.8 Classification of driver alterations

A TRACERx driver gene list was collated using all genes identified in the COSMIC cancer gene census (v75) (Forbes et al., 2015). Additionally, genes found in large-scale pan-cancer analyses (Lawrence et al., 2014), as well as lung cancer specific genes were also included (Cancer Genome Atlas Research Network, 2012, 2014; Campbell et al., 2016). A mutation was classified as a driver mutation, only if it occurred in one of the genes included in the curated driver gene list, and if it fit either of the following criteria:

A) A deleterious mutation (i.e. either a stop-gain mutation, or predicted by two out of three tools to be deleterious – Sift (Kumar, Henikoff and Ng, 2009), Polyphen (Adzhubei, Jordan and Sunyaev, 2013), and MutationTaster (Schwarz et al., 2010)), which was found in a gene annotated as recessive (i.e. tumor suppressor gene) by COSMIC.

B) If the gene was annotated as dominant (i.e., oncogene) by COSMIC and three or more exact matches to the mutation were found within the COSMIC database.
Finally, to classify lung cancer specific drivers, a curated gene list was used including genes reported in (Berger et al., 2016; Jamal-Hanjani et al., 2017; Martincorena et al., 2017; Bailey et al., 2018) as associated with either lung adenocarcinoma or lung squamous cell carcinoma.

### 2.2.2 Processing of pancreatic cohort

#### 2.2.2.1 Sample extraction and whole exome sequencing

For germline DNA extraction, the QIAGEN DNeasy Blood and Tissue kit (Qiagen) was used on peripheral blood (100ul). DNA from tumor tissue was also extracted using the QIAGEN DNeasy Blood and Tissue kit (Qiagen). The germline and tumor tissue samples, including PDX samples were sequenced on the Illumina HiSeq to a median depth of 103x. cfDNA was extracted from 1ml of plasma obtained from sequential blood samples using the QIAamp Circulating Nucleic Acid kit (Qiagen). Exome capture was performed using the SureSelect Human All Exon V5 or V6 kits (Agilent).

#### 2.2.2.2 Patient-derived xenografts

Metastatic samples from the liver, obtained through core needle biopsies, were used to derive patient-derived xenograft (PDX) models for some of the patients' metastatic samples. The tumor cells were implanted subcutaneously into 5 to 6-week-old NOD Scid Gamma (NSG) mice (F1 or engraftment phase) (2-3 mice per tumor). Tumors from these initially engrafted NSG mice were subsequently implanted into 5 to 6-week-old Athymic nude-Foxn1 (nude/nude) mice subcutaneously into both flanks of 5-8 mice (F2 or expansion phase). When the F2 tumors reached approximately 1500mm$^3$, the tumors were harvested and implanted into the experimental mice, to test different lines of treatments. The PDX samples included in the cohort were sequenced from F2 mice with only a single tumor sample per PDX sequenced.

#### 2.2.2.3 Bioinformatics pipeline

The data were processed analogously to the pipeline described above for the TRACERx and PEACE cohorts (see Section 2.2.1). Due to lower coverage of the cohort, the filters to detect mutations were adapted. Additionally, to account for the PDX samples and possible mouse contamination, alignment of PDX samples was adapted. Details of these adaptations are described in Section 6.2.2.
2.2.3 Mutational signature deconvolution

To deconvolve single base substitution (SBS) mutational signatures, the R package deconstructSigs (v1.9.0) (Rosenthal et al., 2016) was used with the COSMIC signatures (v3) from May 2019 as reference (Alexandrov et al., 2020). As the mutational signatures were deconvolved from WES data, the normalization parameter was set to “exome2genome” and only a subset of signatures previously described for the cancer type were included in the deconvolution.

2.2.4 Quantifying selection using dN/dS

As mentioned in Section 1.2.5, the dN/dS method assumes that if no selection is occurring, the rate of nonsynonymous and synonymous mutations should be equivalent given the number of sites where synonymous and nonsynonymous mutations can occur. It can be applied to measure positive and negative selection from sequencing data.

The dN/dS model as described by (Martincorena et al., 2017) builds upon a Poisson framework initially described by (Greenman et al., 2006). In this framework, the number of mutations of a given substitution type $i$ and functional impact $x$ (where $x$ is either synonymous, missense, nonsense or essential splice site) can be modeled as a Poisson process with rate

$$\lambda = t r_i L_{i,x} \omega_x$$

Where $t$ is the density of substitutions, $r_i$ is the relative rate of substitutions $i$ per site, $L_{i,x}$ is the number of sites at which substitution $i$ can occur with functional impact $x$, and $\omega_x$ reflects the effect of selection on the accumulation of mutations for a given functional impact $x$. For synonymous mutations $\omega$ is set to 1 and not inferred by the model, whereas for all other functional impact types, $\omega$ is inferred as the maximum-likelihood estimate using a Poisson regression after correcting for the rates and sequence composition (Martincorena et al., 2017).

The simplest implementation of the dN/dS model only considers a single substitution type for all mutations, while in the model described by (Martincorena et al., 2017) the full trinucleotide model with 192 parameters is considered to account for context-specific mutational processes as well as strand asymmetry. Finally, to account for variations in mutation rates across genes, instead of using a single parameter $t$, it is possible to specify separate mutation rates for each gene. However, this approach
lacks power in smaller datasets. To overcome this issue, (Martincorena et al., 2017) implemented a covariate model which can account for variation in mutation rates due to gene expression, or epigenetic marks with the mutation rate $t$ modelled as Gamma-distributed reflecting the variation across genes.

For the cohort level analysis of selection, an adapted version of dndscv (Martincorena et al., 2017) was used to estimate global dN/dS values for a given gene list. This adapted version first estimates the global rates for the various substitution types using all mutations and genes (equivalent to running the original dndscv function without a gene list). In the original dndscv function, when a gene list is specified, the global rates are estimated only based on mutations in those genes, limiting the confidence as only a small number of mutations are used. In contrast, in the adapted version, the estimates obtained for the global rates of the substitution types are then used for any given gene list specified. This means that comparisons between analyses of various gene lists are possible as the background rates are consistent across all runs.

2.2.5 Inferring migration patterns using MACHINA

MACHINA (El-Kebir, Satas and Raphael, 2018) was developed as a multi-objective optimization problem with additional constraints to infer migration histories in metastatic cancers and quantify occurrences of metastasis-to-metastasis seeding.

Given a clone tree derived from sequencing data as well as a migration pattern, MACHINA attempts to define a labelling of the internal nodes such that first the number of migrations, then co-migrations and finally the number of seeding sites are minimized under a maximum parsimony objective. The migration pattern can be defined as one of the following:

1) Parallel single source seeding (PS): which only allows cell migrations to occur from the primary tumor to each site of metastasis
2) Single source seeding (S): which allows cell migrations from a single source (not necessarily the primary tumor)
3) Multi source seeding (M): which allows cell migrations from multiple sources to a single metastasis
4) Reseeding (R): which allows metastasis 1 to seed metastasis 2 and vice versa.

In a second version of the problem, MACHINA attempts to refine a given clone tree to resolve polytomies and solves the problem above given the refined tree.
Both versions of the migration problem can be solved using an integer linear program (ILP) given a clone tree and specified migration pattern. Additional constraints ensure that the root of the tree is labeled as the primary tumor, that the annotation of the leaves is fixed and that each internal node is only labeled as one anatomical site.
Chapter 3  TRACERx clustering and tree building methods

3.1 Introduction

In order to apply an evolutionary approach to analyzing tumor data, one should ideally sample the tumor multiple times longitudinally to understand how it evolves over time. As the primary treatment for most early-stage tumors is surgery with curative intent where the entire tumor is removed, it is not possible to track its evolution over time. However, by sampling multiple regions of the primary tumor it is possible to reconstruct the evolutionary history of the tumor up to the point of surgery. To do this, one must first group genomic alterations that likely occurred at a similar point in the evolution of the tumor and consequently likely are found in the same cells. Subsequently these clusters can be used to infer the ordering of all events from the initial cancer cell up to the point of sampling.

In this chapter, I will present a novel clustering and tree building method as well as comprehensive benchmarking and validation using a novel simulation framework. The work described is part of a manuscript currently in preparation. The clustering method was adapted from the previously described method in (Jamal-Hanjani et al., 2017) by me, the tree building method was a collaboration between myself, Emma Colliver and Alex Frankell. The simulation framework and benchmarking were developed and performed by Abi Bunkum.

3.2 Main restrictions on tree building

Genomic alterations can be clustered into groups of events that likely occurred at a similar point in the evolution of a tumor. This is often done using the variant allele frequency (VAF) or the derived estimate of the cancer cell fraction (CCF), i.e., the proportion of tumor cells in which such alterations occur (Miller et al., 2014; Roth et al., 2014). In order to calculate the CCF of a given alteration, the VAF is corrected for tumor purity and the copy number at the site. To infer the phylogenetic relationship of such clusters of genomic alterations, two main assumptions are integrated.
Figure 3-1 Cartoon highlighting the pigeonhole principle applied in tree building. Three clusters (light blue, dark blue, and green) with varying cancer cell fractions are shown for a single region. In scenario 1 both subclonal clusters (dark blue and green) have cancer cell fractions exceeding 50%, meaning they necessarily need to be nested within each other. Scenario 2 shows two subclonal clusters where one has a cancer cell fraction below 50%. In this scenario, both a nested and branched tree represent feasible solutions.

The pigeonhole principle (Nik-Zainal, Van Loo, et al., 2012; Jiao et al., 2014) means that in any given region, if two clusters of genomic alterations each have a cancer cell fraction
greater than 50%, one cluster must necessarily arise from the other (Figure 3-1). This is due to the fact that the combined cancer cell fraction of distinct clusters cannot exceed 100% of the tumor cells. For example, considering three clusters with a CCF of 100%, 70%, and 60%, respectively. Cluster 1 is necessarily clonal within the region while clusters 2 and 3 make up a proportion of cancer cells greater than 50% each. Therefore, as the overall cancer cell fraction cannot exceed the CCF of the parental cluster (in this case cluster 1), cluster 2 must be the parental cluster of cluster 3 and therefore the tree structure is a linear relationship of cluster 1 to cluster 2 to cluster 3 (Figure 3-1; Scenario 1). If, however, cluster 3 has a CCF of 30%, both a linear and branched relationship of cluster 2 and 3 would be possible. This means that in this scenario cluster 3 could be nested within cluster 2, but an equally plausible solution would be that both cluster 2 and 3 are descended from the clonal cluster 1 (Figure 3-1; Scenario 2).

The second assumption introduced to infer the phylogenetic tree is the crossing rule (Jiao et al., 2014). This is relevant when multiple regions of a tumor have been sampled. This rule means that if in one region a cluster is nested within another, in a different region of the same tumor the reverse relationship is not possible (Figure 3-2). More specifically, returning to the example above where in region 1 both a linear and branched relationship between cluster 2 and 3 are possible. If in region 2 the CCF of cluster 1, 2 and 3 were 100%, 0% and 40%, respectively, cluster 3 could not be nested within cluster 2. Therefore, the only possible relationship of the two clusters would be a branched event where both clusters are nested within cluster 1 only (Figure 3-2; Scenario 1). If, however, in region 1 the CCFs of cluster 2 and 3 were 70% and 60%, then no possible solution would exist in which both clusters could be assigned to the tree. In this scenario, one of the two clusters would need to be excluded as it breaks the phylogenetic principles (Figure 3-2; Scenario 2). Various approaches can be used to select which of the two clusters should be removed. However, by relaxing the CCF threshold needed to determine the nesting structure by for example 10%, both subclonal clusters could be retained.
Three clusters (light blue, dark blue, and green) with varying cancer cell fractions are shown for two regions (R1 and R2). In scenario 1, both a branched and nested tree are feasible using the cluster estimates of R1, however as in R2 cluster 2 (dark blue) is absent, cluster 3 (green) cannot be nested within cluster 2, therefore only a branched tree represents a feasible solution when taking into account both regions. In scenario 2, for R1 cluster 3 is nested within cluster 2, while in R2 cluster 2 is absent and therefore cluster 3 cannot be nested within cluster 2. In this case, no plausible tree solution exists without removing one of the two clusters, therefore the two possible tree solutions represent linear trees with one of the two subclonal clusters removed, respectively. In the second scenario, by relaxing the CCF thresholds and including a CCF buffer (for example of 10%), it is possible to allow cluster 2 to be nested within cluster 3 in R1. If such a relaxation is implemented, a tree using all three clusters can be constructed.
3.3 Overview of existing methods

Due to the heterogeneity within not only the tumor, but also each sampled region, state-of-the-art phylogenetic methods cannot be directly applied to analyze tumor evolution. Previously, studies have inferred phylogenetic trees from multi-region sampling at a regional level using maximum parsimony or neighbor joining (Kim et al., 2015; Gibson et al., 2016; Zhao et al., 2016; Brown et al., 2017; Naxerova et al., 2017; Zhai et al., 2017; Reiter et al., 2020; Tang et al., 2021). These approaches implicitly assume that each region contains a homogeneous population of cells and therefore under-estimate the heterogeneity within a tumor. Furthermore, frequently the most parsimonious solution will lead to erroneous inference of tumor evolution including convergent evolution and homoplasies (Alves, Prieto and Posada, 2017) (the same mutation occurring on multiple clades of the phylogeny) which is a direct violation of the infinite sites assumption (Kimura, 1969).

Instead, an initial deconvolution of the different cell populations present at the point of sampling is necessary to then infer the evolutionary history of a tumor. Multiple such methods exist for single nucleotide variants (SNVs) and insertions and deletions (INDELs). Some methods just cluster mutations (Miller et al., 2014; Roth et al., 2014), others reconstruct the phylogenetic relationships (Malikic et al., 2015; Popic et al., 2015; Dang et al., 2017; Satas and Raphael, 2017), while others again will both cluster mutations and infer phylogenetic trees (Deshwar et al., 2015; El-Kebir et al., 2015, 2016; Malikic et al., 2015; Popic et al., 2015; Yuan et al., 2015; Dang et al., 2017; Wintersinger et al., 2022). To date, these methods can only cluster and infer the relative timing of mutations. No comparable methods exist for copy number aberrations where phylogenetic reconstruction is only possible at a regional level so far.

3.4 Previous implementation in TRACERx100

In the previous analyses of TRACERx data (Jamal-Hanjani et al., 2017), a modified version of PyClone (Roth et al., 2014) was used to cluster mutations. Initially, the observed mutation copy number was calculated from the variant allele frequencies by also taking into account tumor purity and local copy number. Subsequently, mutations were corrected for subclonal copy number to avoid overestimating the number of subclonal mutations that could be attributed to subclonal copy number events. For example, if a mutation was present in regions 1 and 2 but absent from regions 3, if there was a copy number loss at the position of the mutation, it could be assumed that the
Chapter 3 Results

mutation was previously found in all three regions, but subsequently lost in region 3. Therefore, this mutation was corrected and merged with the clonal mutations present in all three regions. These corrected cancer cell fraction estimates were used as input to PyClone (v0.12) by including the observed variant counts but adjusting the reference counts such that the VAF would equal half the estimated CCF. As the input was already corrected for purity and ploidy of each sample, the major allele copy number was set to 2 and the minor allele copy number was set to 0, and the purity of each sample was set to 0.5.

Subsequently, a manual phylogenetic reconstruction was attempted, removing clusters of mutations that were likely driven by copy number events, clusters with less than 5 mutations, as well as any clusters that broke the two main hypotheses used to infer the phylogenetic trees (see Section 3.2). Finally, CITUP (v0.1.0) (Malikic et al., 2015) was run on the manually curated set of mutation clusters to verify the manually inferred phylogenetic trees.

3.5 Motivation for updating method

Initially, the version of the clustering as described in (Jamal-Hanjani et al., 2017) was applied to the TRACERx paired primary metastasis as well as the PEACE cohort. Due to the large number of samples per patient within PEACE, the clustering approach did not converge and did not provide any output and therefore could not be applied to many of the tumors.

Additionally, when applying the clustering method to the paired primary metastasis cohort and inferring the dissemination patterns using the mutation clusters, the proportion of polyclonal dissemination was much higher than when simply using the presence or absence of mutations. This was mainly driven by clusters present at low frequency either in the primary tumor or metastasis samples, where the majority of mutations were absent in the given region, but a few mutations were driving the mean CCF to greater than 0. This meant that overall, the cluster was defined as present in a given region, and therefore multiple clusters were shared between the primary tumor and metastasis. By introducing a minimum proportion of mutations with CCF greater than 0 to define a cluster as present in a given region, this bias was rectified, and the proportion
of polyclonal dissemination was reduced and comparable to the naïve method of using presence/absence of mutations.

![Figure 3-3 Comparison of dissemination patterns using the previous clustering and updated clustering approaches. The number of metastases samples with either monoclonal or polyclonal dissemination are shown for the previous clustering approach, as described in (Jamal-Hanjani et al., 2017) (on the left) compared to the updated clustering approach described below (on the right).](image)

Therefore, the clustering approach from (Jamal-Hanjani et al., 2017) was adapted further to reduce bias in defining polyclonal metastases due to misclassifying presence/absence of clusters (see a comparison of the two methods in Figure 3-3). Additionally, a parallelization was introduced to allow clustering of mutations in tumors with large number of samples or mutations.

### 3.6 Updated approach

#### 3.6.1 Mutation clustering

An updated version of PyClone (v0.13.1) (Roth et al., 2014) was used and modified to cluster mutations and classify them as clonal or subclonal. As mentioned above, to
calculate the phylogenetic CCF (phyloCCF), initially the VAF, tumor purity and local copy number were used to calculate the expected mutation copy number. Subsequently, subclonal copy number correction was performed as described previously (Jamal-Hanjani et al., 2017). Additionally, to avoid mis-clustering of INDELs into INDEL driven clusters, a region specific VAF correction factor was applied (Jamal-Hanjani et al., 2017).

As mentioned above, one of the main issues with the previous implementation of the clustering of mutations was the over-estimation of clusters present in any given region. Secondly, the large number of regions sampled for some tumors, especially in PEACE meant that the clustering did not complete successfully. Therefore, a pre-clustering step was implemented based on binary presence/absence of mutations and subsequently, PyClone could be run on these individual clusters using solely the regions the clusters were found in. This was possible as including any regions in which the group of mutations was absent did not introduce any additional information to the clustering step. When considering a group of mutations present in regions 1 and 2 and absent from regions 3-5, there would be no additional value added by including regions 3-5 in the clustering and indeed by adding these regions, more noise might be introduced in the clustering of the mutations. Furthermore, mutations occurring in different regions of the tumor must necessarily have occurred on separate branches of the phylogenetic tree (see Section 3.2) and therefore no additional information could be gained by clustering these groups of mutations simultaneously. Once again, considering a group of mutations present in regions 1 and 2 and another group of mutations present in region 5, these mutations necessarily would cluster in at least two separate clusters, or otherwise unintentional noise might have been added from low CCF mutations. Therefore, PyClone could be run on each pre-clustered set of mutations separately, in parallel, speeding up the clustering process significantly.
In the first step, the presence or absence of each mutation across each region is inferred. Subsequently, mutations occurring in the same regions are grouped together to create pre-clusters. Any regions where a set of mutations does not occur can be ignored for the remaining steps. In the next step, the cancer cell fraction (CCF) of each mutation is calculated and subsequently each pre-cluster is clustered using PyClone based on the CCF estimates to form additional clusters.

More specifically, these pre-clusters were created by taking into account the presence or absence of each mutation in each region. A mutation was classified as present if there was at least one read supporting the variant as well as a VAF greater than 1%. Subsequently, the mutations were grouped together based on which regions they were present in (Figure 3-4; left panel). If less than 5 mutations were grouped together, this mutation-set was not clustered using PyClone.

All other mutation-sets were then independently clustered using the PyClone Dirichlet process clustering (v0.13.1) (Roth et al., 2014) only using the individual pre-clusters and the regions these appeared in as input (Figure 3-4; right panel). The data used as input consisted of the observed variant read count, as well as the reference count which was adjusted such that the VAF was equal to half of the calculated phyloCCF. Major allele copy number was set to 2 and minor allele copy number was set to 0, with a purity of 0.5 also set as input. This allowed clustering of mutations based on the previously calculated phyloCCF.

PyClone was then run with 10,000 iterations and a burn-in of 1000. Additionally, the reference prior was set to normal, and the variant prior was set to BB only, which essentially assumes that each mutation is diploid and homozygous (Roth et al., 2014). Furthermore, the maximum number of clusters was limited to avoid over-clustering and
set so that at least five mutations could be present in each cluster, or a maximum of 10 clusters overall. All other parameters were left as their default values.

The clustering step was run on all regions of the primary tumor, as well as any metastatic regions or regions from tissue harvest, if available for a patient. By combining all possible regions, the resolution of the clustering is improved by separating mutation clusters that might have otherwise been grouped as one if less regions were considered. Therefore, the cluster defined as truncal was clonal in both the primary and metastases, while all clusters defined as subclonal could be shared, or primary or metastasis unique. Furthermore, such subclonal clusters could still be defined as clonal in individual regions, but not overall within the tumor.

After the clustering was performed using PyClone, clusters were investigated to identify occurrences where copy number loss events might have caused low phyloCCF estimates. This was done in the same way as was described previously in (Jamal-Hanjani et al., 2017). Furthermore, to detect whether clusters might have been driven by undetected subclonal copy number events, the genomic distribution of mutations from each cluster was considered. For each subclonal cluster, the number of chromosomes where mutations of each cluster were observed was calculated. A background distribution was calculated using a randomly resampled set of clonal cluster mutations and this was repeated 10,000 times. The likelihood for the observed cluster distribution was compared to the background distribution and a p-value was calculated. If the number of chromosomes of the observed cluster was significantly lower than the background distribution (p = 0.01), such clusters were labelled as “copy-number driven clusters”.

To avoid incorrect clusters which were driven by the subclonal copy number correction performed earlier, each cluster was tested as to whether it could be merged with other clusters when the copy number correction was removed. Finally, to avoid overestimating heterogeneity, clonality calls from the clustering were reviewed and any cluster for which more than 25% of mutations were found at a phyloCCF of 90% or greater in all tumor regions were re-defined as clonal and merged with the previously defined clonal cluster.

3.6.2 Phylogenetic tree building

To construct phylogenetic trees from clustered mutations in an unbiased and automated way, a computational framework was created which applies the pigeonhole principle and
crossing rule (Jiao et al., 2014) and subsequently removes any subclonal clusters that violate these assumptions, optimizing for the number of mutations removed.

The clusters identified using the approach set out above were used as input to the tree building, removing any clusters with less than 5 mutations. For all other clusters, a nesting matrix was obtained which described the ancestral-descendant relationship between clusters. For this, each tumor region was considered separately and the CCF distributions for every pair of clusters was compared using a Wilcoxon rank-sum test (p = 0.01). For example, cluster 1 could potentially be nested within cluster 2 if the phyloCCF of each mutation of cluster 1 was significantly lower than the phyloCCFs of cluster 2 across each region, using a p-value threshold of 0.01. Through the nesting matrix the truncal cluster was identified as the cluster which could not be nested in any other and that all other clusters could be nested within.

Subsequently, a directed graph was constructed using the R function “graph.data.frame” and the nesting matrix. Issues in the tree structure were evaluated based on circles in the graph, where the same cluster was descended from two separate clusters, indicating a violation of the crossing rule; and whether the CCF of any level exceeded 100% (with an added noise buffer of 10%) which indicated a violation of the pigeonhole principle.

![Figure 3-5 Example of output from the tree building method.](image)

On the left, the cancer cell fractions (CCF) for each cluster and region are shown as a bar plot and the meanCCF is also indicated for each cluster. Individual CCF estimates of each mutation within a cluster are
overlayed as a scatterplot and the bootstrapped confidence intervals are also shown in black. On the right, the tree solution is shown. The gray pie chart is split into $n$ segments representing each of the regions included. On the tree, each of the clusters is represented as a pie chart with the same orientation as the gray pie chart in the corner and colored segments highlight presence of the cluster in a given region, with lighter colors representing lower CCF values. Grayed out segments represent absence of a given cluster in that region. Edges colored in black represent consensus branches, i.e., the same parent-child relationship can be found in all alternate trees, while for edges in gray, the parent-child relationship differs across the different tree solutions. Finally, in the top right corner the number of mutations associated with each cluster is also shown.

To resolve any cases with biologically implausible solutions, clusters were removed in an iterative manner until a biologically meaningful tree could be found. This was optimized for the number of mutations retained on the tree. Larger clusters could be estimated with greater confidence compared to smaller clusters which could appear due to sequencing noise of missed subclonal copy number alterations. Therefore, clusters with more mutations were prioritized and clusters were iteratively removed to maximize the number of mutations retained.

Once a tree was identified for a given tumor (see example output in Figure 3-5), all other alternate trees using the same set of clusters were evaluated. Once all alternate trees were enumerated, a set of consensus branches was determined as the conserved parent-child relationships.

3.6.3 Classifying clonality and presence of individual clusters in tumor regions

Each mutation cluster was defined as absent or present in a given region using the phyloCCF of the mutations and the clusters defined as present were further split into clonal and subclonal.

If the mean phyloCCF of a cluster was 0 in a region, this cluster was defined as absent from that region. To determine whether a cluster was clonal or subclonal in a given region, the phyloCCF of the mutations was compared to the phyloCCFs of the truncal cluster within that region, instead of using a fixed threshold. In general, the truncal cluster usually had a mean CCF close to 100%, however in some cases due to low purity or slightly incorrect copy number estimation, the mean CCF of the truncal cluster could be higher or lower than 100%, meaning that all other CCF estimates might have presented
the same bias. Therefore, using a fixed threshold might have led to incorrect classification of a cluster as clonal or subclonal.

Consequently, the distribution of phyloCCF of mutations for each cluster of interest were compared to the phyloCCFs of the truncal cluster using a one-sided Wilcoxon test. If no significant difference was observed ($p \geq 0.05$) between the cluster of interest and the truncal cluster, the cluster of interest was defined as clonal within that region. Furthermore, a lower threshold of the truncal cluster was calculated by using the lower bound of the 95% bootstrapped confidence interval of the truncal cluster or 0.9. If the upper bound of the 95% confidence interval of the cluster of interest overlapped with the lower threshold of the truncal cluster, the cluster of interest was also defined as clonal. Conversely, if the distribution of phyloCCFs of the cluster of interest was significantly lower than the phyloCCFs of the truncal cluster (Wilcoxon one-sided test, $p < 0.05$) and the upper bound of the cluster of interest did not overlap with the lower threshold of the truncal cluster, the cluster was classified as subclonal.

Finally, based on the classifications of each cluster in each region, these could be extended to tumor level definitions, e.g., clonality of a given cluster in the primary tumor, all metastatic sites, individual tissue types, or for each metastatic tumor separately. Clusters that were clonal across all regions of interest (i.e., all primary tumor regions, or all metastatic samples) were defined as clonal overall within the primary tumor or across metastatic samples, respectively. Clusters that were subclonal or absent from at least one of the regions of interest were defined as subclonal overall, while clusters that were absent from all regions of interest were defined as absent overall.

### 3.7 Benchmarking and validating approach

#### 3.7.1 Simulation framework for tumor evolution

To validate the performance of the TRACERx methods, a tumor evolution-informed simulation framework was created. This framework was used to generate realistic tumor phylogenies and simulated bulk sequencing data by matching the specific features found in the sequencing data of the TRACERx 421 cohort, such as tumor purity, mutation burden, number of samples sequenced, and sequencing coverage. The simulation framework models the evolution of different types of genetic alterations that are pervasive in non-small cell lung cancer (NSCLC), including single nucleotide variants...
(SNVs) and somatic copy number aberrations (SCNAs) (both gains and losses) as well as clonal and subclonal whole genome doublings (WGDs). Importantly, the framework specifically models the interactions and overlap of these events, in particular SCNAs affect the mutation multiplicities of overlapping SNVs and deleterious SCNAs can result in mutation losses.

Phylogenetic tree topologies were simulated with \( n \) = 8-16, 12-24, 22-30 clones for low (2-3 samples), medium (4-7 samples), and high (>7 samples) sample groups, respectively. For this, a binary tree with \( n \) leaves was generated and random leaves were removed iteratively until the tree contained a total of \( n \) nodes (both ancestral and leaf nodes).

To simulate the evolution of somatic mutations and genomic alterations, the number of mutations in the most recent common ancestor (MRCA) and all other nodes of the tree were sampled from the empirical distribution based on the TRACERx 421 cohort. SNVs were randomly assigned to the edges of the tree, preserving the proportion of truncal mutations as <90% of the total mutation burden. Subsequently, copy number gains and losses were also assigned to the nodes based on the empirical distributions of SCNAs. Each copy number event can affect the mutation multiplicity when such an event is assigned to the same allele as the mutation (which is assumed to always occur on allele 1, without loss of generality). Importantly, the simulation framework follows two main assumptions: (1) the infinite sites assumption which states that each SNV occurs only once during tumor evolution, and (2) the constant mutation multiplicity assumption, such that every SNV has the same mutation multiplicity across different tumor clones.

Bulk tumor samples are obtained from the simulation framework by sampling both the normal diploid clone as well as \( m \) distinct tumor clones, with \( m \) chosen uniformly at random from between 3 and 8 clones, as observed in previous studies (Dentro et al., 2021). The proportion of normal diploid cells to tumor cells is determined by the tumor purity which is sampled from the provided empirical distribution. Equally, the overall number of tumor regions is sampled from the provided empirical distribution. Finally, simulated DNA sequencing data is derived for every bulk tumor sample, including the total number of reads and the observed number of variant reads. This simulated DNA sequencing data then is used as input to the benchmarking described below.
3.7.2 Benchmarking different clustering and tree building approaches

Based on these simulations, the performance of the novel clustering and tree building methods were evaluated and compared to other state-of-the-art approaches including PyClone (Roth et al., 2014) (clustering only), phyloWGS (Deshwar et al., 2015), LICHeE (Popic et al., 2015), and CITUP (Malikic et al., 2015) (clustering and tree building). In the following only the benchmarking for the clustering methods will be presented.

The results demonstrate that the TRACERx methods for clustering and tree building, described above, outperform the other approaches, especially in three key areas which are particularly relevant in performing key analyses described in this thesis. These areas are: (1) accurate reconstruction of tumor evolutionary histories in the presence of...
complex NSCLC evolutionary events, including clonal and subclonal WGDs and mutation losses, (2) robust classification of the evolutionary timing of somatic mutations (truncal vs subclonal), and (3) confident classification of presence or absence of mutations in a given region.

The approaches were compared on a dataset of 150 simulations, as described above, with three different categories of tumor samples: 2-3 samples for the low category, 4-7 samples for the medium category, >7 samples for the high category. Most importantly, the clustering method successfully completed within a reasonable time frame of 8 hours for all simulations tested, even for larger sample sizes (Figure 3-6). Other methods tested did not always complete in the time frame set. This was especially true for phyloWGS where no runs completed within 8 hours for the medium and high sample groups and only three runs completed successfully in the low sample group (Figure 3-6). Therefore, it was not included in additional comparisons.

![Figure 3-7 Mutation clustering ARI for the different methods.](image)

*The three main panels show the different sample sizes that were considered with each panel encompassing 50 simulations. The left-hand panel shows the sample group low which includes 2-3 regions, the middle panel shows sample group medium with 4-7 regions, while the right-hand panel shows sample group high with more than 7 regions. As phyloWGS timed out for the majority of simulations, it was excluded from this*
The mutation clustering adjusted rand index (ARI) was compared across the four remaining methods (TRACERx clustering, PyClone, LICHeE, and CITUP) (Figure 3-7). Across all the sample groups, the TRACERx clustering method outperformed the other methods, with LICHeE performing worst. Interestingly, with increasing sample sizes, the mutation clustering ARI of most methods improved, although for PyClone the median stayed consistent across sample groups (Mutation clustering ARI for PyClone: low: 0.743, medium: 0.712, high: 0.739).

The three main panels show the different sample sizes that were considered with each panel encompassing 50 simulations. The left-hand panel shows the sample group low which includes 2-3 regions, the middle panel shows sample group medium with 4-7 regions, while the right-hand panel shows sample group high with more than 7 regions. As phyloWGS timed out for the majority of simulations, it was excluded from this analysis, therefore for each sample group only 4 methods are shown. The mutation presence precision is shown for each method and sample group, which describes how accurately mutations were defined as present or absent in a given region.
Chapter 3 Results

Additionally, the methods were compared on how accurately mutations were defined as present or absent (mutation presence precision) (Figure 3-8). Here, all methods performed relatively well for low sample numbers, with CITUP, and to a lesser extent PyClone, performing progressively worse for larger sample groups. LICHeE and the TRACERx clustering method both had high mutation presence precision across all sample groups, however LICHeE failed in 15 of 50 mutations in the high sample group, which could bias the results seen.

3.8 Conclusions

The results in this chapter highlight that especially for greater sample size, the novel methods described here outperform other state-of-the-art methods used previously. Due to the pre-clustering step, which was implemented, the actual clustering step can be performed in parallel, reducing computation time significantly. Therefore, the TRACERx methods described here completed successfully in less than 8 hours for all simulations across various sample sizes, while other methods, such as CITUP (Malikic et al., 2015) and especially phyloWGS (Deshwar et al., 2015) failed to complete in the set time frame.

The output from the clustering of the TRACERx method had a higher mutation clustering ARI compared to the other methods, highlighting that the clusters were more similar to the ground-truth clusters from the simulations. Furthermore, the TRACERx method was also better at determining presence or absence of mutations compared to the published methods. This is particularly important when estimating dissemination patterns, as it reduces artificial heterogeneity by removing clusters driven by a small subset of mutations which could just be due to sequencing artefact or noise.

The methods presented here will be used in subsequent chapters to infer the phylogenetic relationships of primary and metastatic samples. These phylogenies will then be leveraged to better understand evolutionary trajectories in metastases by defining the dissemination patterns from primary tumor to metastasis, as well as inferring the timing of metastatic divergence.
Chapter 4  Characterizing the seeding clone in metastasis

4.1 Introduction

One key unanswered question in cancer biology is why tumors metastasize. Although tumor stage is a good predictor of outcome, some patients with early-stage tumors will develop recurrence of their disease, while other patients with larger tumors and nodal involvement do not experience a relapse of their disease. To better understand the metastatic potential of primary tumors we first must understand the events which engender it. This may help elucidate whether metastatic vulnerabilities exist which could be targeted by existing or new therapeutic approaches (Brastianos et al., 2015; Esposito, Ganesan and Kang, 2021).

By characterizing which cell populations are shared between the primary tumor and metastases, it is possible to investigate the underlying drivers of metastatic potential. Each metastatic cell will include all genomic alterations accumulated during the evolution of the primary tumor, up to the point of metastatic divergence, as well as any additional mutations arising within the metastasis. Previous studies exploring single sample primary or metastasis tumors have not identified many metastasis-unique cancer genes, although some differences between primary and metastatic samples were observed (Priestley et al., 2019; Nguyen et al., 2022). However, it is unclear whether genomic alterations observed in metastatic samples were metastasis specific, or present in the paired primary tumor and enriched in the metastatic samples. Therefore, to fully explore drivers of metastatic potential, paired primary-metastatic cohorts, such as the one presented here are needed.

By leveraging the paired primary-metastatic data it was possible to define metastasis-unique and primary-unique mutations as well as mutations shared between the primary tumor and metastasis, of which a subset could be defined as metastasis seeding. Subsequently, mutations occurring in the same genes, within the seeding mutations were explored and compared to events occurring solely in the primary tumor. These observations could give an indication on recurring patterns within the primary tumors
which could drive metastatic potential and therefore allow prediction of when metastases are likely to arise.

In the following chapter, I will introduce metastatic dissemination from the primary tumor and explore these definitions in the TRACERx paired primary metastasis cohort. Additionally, by utilizing the TRACERx clustering and tree building method, I will define the seeding clone within the primary tumor and explore selection which could lead to metastatic potential.

The work described in the following chapter is part of a manuscript currently under review. The TRACERx paired primary metastasis data was collected by members of the TRACERx consortium, which I am also part of, and processed by me through a pipeline developed and adapted by members of the TRACERx consortium, including myself. All analyses described in the following section were conducted by me, if not otherwise stated.

4.2 Modes of metastatic dissemination

Metastatic potential may be acquired once or multiple times during tumor development. If metastatic potential is acquired early in tumor development, the trait is likely shared across all cells of the primary tumor which could potentially result in polyclonal and polyphyletic seeding patterns of metastases (Notta et al., 2016; R. Gao et al., 2016). On the other hand, monoclonal dissemination likely occurs if the metastatic potential is acquired later in tumor evolution and in only one or very few subclones (Turajlic, Xu, Litchfield, Rowan, Chambers, et al., 2018). Knowing the metastatic dissemination patterns could indicate when combination therapy is likely to be required, as might be predicted in cases of polyclonal and polyphyletic dissemination.

4.2.1 Defining dissemination patterns from the primary tumor based on shared clonality

To understand metastatic potential and patterns of metastasis it is critical to have a clear framework to define patterns of metastatic dissemination. Dissemination from the primary tumor can be explored separately for each metastatic sample (sample level), or by combining all metastatic samples of a given tumor to define overall (tumor level) dissemination patterns.
4.2.1.1 Sample level dissemination patterns based on shared clonality

Monoclonal dissemination from the primary tumor can be defined as a single cell population from the primary seeding the metastatic sample (Figure 4-1). In this case, a bottlenecking event occurs such that all mutations that are shared between the primary tumor and metastatic sample appear clonal within the latter. If any shared mutations appear subclonal in the metastatic sample, this necessarily means that multiple clones from the primary tumor must have seeded the metastasis. We refer to this as polyclonal dissemination from the primary tumor (Figure 4-1).

In both cases of dissemination, there can be ongoing evolution in the metastasis leading to additional clones being present, however for the purpose of these definitions only the shared clones between the primary tumor and metastasis are of interest.

![Figure 4-1 Schematic of metastatic dissemination from the primary tumor for individual metastatic samples. On the left, the composition of the primary tumor is shown and a possible composition of the metastatic sample in the case of monoclonal dissemination. This highlights that only one clone is shared between the primary tumor and the metastatic sample. On the right-hand side, an example of the composition of a metastatic sample seeded polyclonally from the primary tumor is shown, highlighting that multiple clones are shared between the primary tumor and the metastatic sample.]

4.2.1.2 Tumor level dissemination patterns based on shared clonality

To define dissemination patterns at the tumor level, firstly the mode of dissemination for all samples was considered. Subsequently, the cancer cells shared between the primary tumor and all metastatic samples were considered. If all metastatic samples of a given tumor showed patterns of monoclonal dissemination and the cancer cell populations shared between the primary tumor and the metastatic samples were the same across all metastatic samples, in the following this is referred to as uniform monoclonal dissemination (Figure 4-2). On the other hand, if different metastatic samples shared...
distinct cell populations with the primary tumor, this is referred to as polyclonal dissemination. Polyclonal dissemination can be split further into tumors where all metastatic samples were seeded by a single clone in the primary tumor and those where at least one metastatic sample was defined as polyclonal dissemination. The former of these cases can be referred to as parallel monoclonal dissemination while the latter can be referred to as polyclonal dissemination in the following (Figure 4-2).

If we consider the evolutionary history of the disseminating cancer cells, uniform monoclonal dissemination was necessarily monophyletic in origin, as only one clone of the primary tumor was driving metastatic dissemination (Figure 4-2). Therefore, necessarily only one branch of the phylogenetic tree could be involved. In the cases of parallel monoclonal or polyclonal dissemination it was possible to increase the resolution on the patterns of dissemination further by taking into account the phylogenetic tree and the relationships of the clones involved in metastatic dissemination. If all seeding clones could be mapped to a single branch of the tree, the overall pattern was referred to as monophyletic in origin (Figure 4-2). If the seeding clones lie on distinct branches of the evolutionary tree, the overall pattern was referred to as polyphyletic (Figure 4-2).

Tumors with monophyletic dissemination likely acquired their metastatic potential once in their evolutionary history. This likely occurred later on in tumor evolution, once the cell populations had branched, as otherwise additional cell populations might have shared this metastatic potential. Alternatively, the event might have occurred earlier in tumor development with additional selection pressures acting such that only this single subclone could metastasize, for example multiple cell populations metastasized, however one subpopulation had a fitness advantage at the metastatic site outcompeting other cell populations.

Conversely, in tumors with polyphyletic dissemination the metastatic potential was likely acquired early in tumor evolution and shared across all subpopulations. If the metastatic potential was acquired later, multiple such events must have occurred in parallel leading to distinct branches of the phylogenetic tree seeding metastases.
4.2.2 Metastatic dissemination patterns in TRACERx paired primary metastasis cohort

The definitions based on shared clonality (described above) were used to determine the dissemination patterns for each individual metastatic sample within the TRACERx paired primary metastasis cohort, and for tumors with more than one metastatic sample, the tumor level dissemination patterns were also inferred.

In total, 227 metastatic samples from 129 patients were sequenced and processed through the bioinformatics pipeline highlighted in Section 2.2.1, including clustering and tree building (Chapter 3). When determining the dissemination patterns for each metastatic sample independently using the patterns of shared clonal mutations between primary and metastatic samples, the predominant pattern observed was monoclonal dissemination from the primary tumor (183 metastasis samples, 80% overall) (Figure 4-3 A). This suggests that most tumors have a single clone that can metastasize to a specific site, or that additional clonal sweeps are likely to occur within the metastasis.

Furthermore, a slight enrichment of polyclonal lymph node metastases could be observed compared to recurrence and progression samples which were more likely
monoclonal (Fisher's exact test, $p = 0.032$; Figure 4-3 B). This is consistent with previous results (Reiter et al., 2020).

![Dissemination patterns per metastatic sample.](image)

The left plot shows the dissemination patterns for individual metastatic samples. The plot on the right shows the dissemination patterns at a metastasis sample level split by lymph node metastases resected at surgery (labelled primary LN above) and subsequent later timepoint metastatic samples (labelled recurrence above), such as samples from initial recurrence of the tumor or subsequent progression of metastatic disease.

When determining dissemination patterns at the tumor level, the predominant pattern was still monoclonal (85 tumors, 66%; Figure 4-4). Thus, metastatic potential was likely acquired once during tumor evolution, or the detected clone that seeded metastases was fitter than all other clones with metastatic potential.
In the remaining tumors (44 tumors, 34%), multiple clones from the primary tumor gave rise to metastases. In some cases, this could be attributed to tumors with multiple metastatic samples where all samples individually occurred due to monoclonal dissemination from the primary tumor. However, different cell populations seeded different metastatic samples (parallel monoclonal dissemination). Out of the 44 tumors with polyclonal dissemination patterns, 11 (25%) tumors had multiple metastatic samples where all samples were consistent with monoclonal dissemination from the primary and therefore these tumors could be classified as following a parallel monoclonal dissemination pattern. The remaining 33 (75%) tumors had at least one metastatic sample which was defined as polyclonal dissemination from the primary and therefore the overall tumor-level dissemination pattern was defined as polyclonal (Figure 4-4, Figure 4-5).

**Figure 4-4 Overview of dissemination patterns in TRACERx paired primary metastasis cohort.**
Figure 4-5 Dissemination patterns for each metastatic sample grouped by tumor.
The tumors on the left-hand side are classified as monoclonal dissemination while the tumors on the right-hand side are classified as polyclonal dissemination overall. The bar plot shows all metastatic samples arising from each individual tumor and their corresponding dissemination type.

Finally, by taking into account the tree structure, the evolutionary origin of the metastatic clones could be inferred. As mentioned above, tumors defined as uniform monoclonal dissemination were necessarily classified as monophyletic, as by definition the seeding clone originated from a single branch of the phylogenetic tree. If more than one cell population seeded various metastatic samples, it was determined whether these clones originated from a single branch or multiple branches of the tree.

Of the 44 tumors classified as polyclonal dissemination from the primary tumor, 27 (61%) were monophyletic in origin while 17 (39%) were polyphyletic in origin. This highlights that although polyclonal dissemination was more predominant when considering tumor level classifications compared to metastasis level classifications, monophyletic origin was the most predominant dissemination pattern (112 of 129 tumors, 87%). Therefore, even though multiple clones might harbor metastatic potential this event still likely occurred once along a single branch of the phylogenetic tree.
4.2.3 Sampling bias due to number of metastasis regions sampled

As mentioned above, the overall dissemination pattern from the primary tumor is defined by taking into account the dissemination patterns of each metastatic sample. Therefore, the more metastatic samples exist for a single tumor, the more likely it is that at least one of these samples is defined as polyclonal dissemination.

When comparing the number of metastatic samples between cases defined as monoclonal versus polyclonal dissemination, the number of metastasis regions in the cases with polyclonal dissemination were significantly higher than in the cases with monoclonal dissemination (Wilcoxon rank sum test, \( p = 0.0017 \); Figure 4-6). In fact, the more metastatic regions were sampled the higher the proportion of cases with polyclonal compared to monoclonal dissemination (Figure 4-6). In tumors where only 1-2 metastatic samples were analyzed, the predominant dissemination pattern observed was monoclonal (77 of 106 tumors, 72.6%). Conversely, when considering tumors with 5-6 metastatic samples analyzed, the proportion shifted and polyclonal dissemination was the predominant pattern observed (4 of 6 tumors, 66.67%).

This strongly suggests that under-sampling of the metastases can result in misclassification of the dissemination patterns associated with the primary tumor. Therefore, in the clinical setting, where sampling multiple metastatic sites is often not possible due to patient frailty or difficulty in accessing the site of disease, the cell populations with metastatic potential might be insufficiently identified.
4.2.4 Determining the seeding clone

By determining the seeding clone within the primary tumor, i.e., the population of cancer cells that metastasize, it may be possible to better understand the metastatic process and the characteristics of the cancer cells that metastasize.

The seeding clone is defined as the last shared clone between the primary tumor and the metastasis, or the most recent common ancestor between the metastatic sample and any region of the primary tumor. If the dissemination pattern is defined as monoclonal, only one seeding clone exists. However, if the dissemination pattern is polyclonal, multiple seeding clones necessarily exist. Seeding clones were initially defined for each individual metastatic sample. To define the seeding clones found in a given tumor, all seeding clones of each metastatic sample were combined. In the case of polyclonal but monophyletic dissemination, in addition to the most recent common ancestor between...
the primary tumor and metastasis, any other clone present in the metastasis was also considered as a seeding clone.

**4.2.5 Comparison of dissemination patterns to MACHINA**

To ensure confidence in the definitions of dissemination described above as well as the methods to determine the dissemination patterns and seeding clones, an independent tool, MACHINA (El-Kebir, Satas and Raphael, 2018), was used for comparison. MACHINA was developed to infer migration histories in metastatic cancers and quantify occurrences of metastasis-to-metastasis seeding.

It can be run in different modes:

1) Combined tree inference and migration pattern inference: in this mode, MACHINA infers both the phylogenetic tree and migration patterns together by taking as input the frequencies of each mutation and associated confidence intervals

2) Migration pattern inference: in this mode, MACHINA takes a clone tree as input (see Section 5.4.1) and only infers the migration patterns.

3) Migration pattern with tree resolution: in this mode, MACHINA takes a clone tree as input and infers migration patterns while also resolving any occurring polytomies on the clone tree.

Additionally, MACHINA can be run using various hypotheses:

1) Parallel single source seeding – or primary seeding (PS): this hypothesis only allows cell migrations to occur from the primary tumor to all metastases

2) Single source seeding (S): this hypothesis allows cell migrations from one source to one metastasis, where the source can also be another metastasis. Therefore, under this hypothesis metastasis-to-metastasis seeding can be explored.

3) Multi source seeding (M): this hypothesis allows cell migrations from multiple sources to a single metastasis to occur.

4) Reseeding (R): under this hypothesis, metastasis 1 can seed metastasis 2, which in turn can (re-)seed metastasis 1.

While MACHINA is able to infer monoclonal versus polyclonal seeding patterns, it does not take into account the evolutionary tree further, therefore distinguishing between monophyletic and polyphyletic dissemination is not possible. More specifically, for each single metastatic sample MACHINA infers whether a single clone or multiple clones seed
the metastasis. However, when multiple metastatic samples exist, MACHINA does not distinguish whether multiple clones are involved in seeding the different metastatic samples. Consequently, an individual metastasis is defined as polyclonal by MACHINA if it shares multiple clones with the primary tumor (equivalent to polyclonal with monophyletic or polyphyletic origin described above). A metastasis is defined as monoclonal by MACHINA if it shares a single clone with the primary tumor (equivalent to uniform monoclonal, or parallel monoclonal with monophyletic or polyphyletic origin described above). If multiple metastatic samples exist for a given tumor and if any sample is defined as polyclonal, the overall pattern is defined as polyclonal. Conversely, if all individual samples are defined as monoclonal by MACHINA, the overall pattern is also defined as monoclonal.

The definitions based on shared clonality are able to differentiate this latter case further: into whether all individual samples share the same clone with the primary tumor, or if multiple clones from the primary tumor give rise to the individual samples. Therefore, to be able to compare the results from MACHINA and the method described above, the distinction between monophyletic and polyphyletic dissemination was ignored, and uniform monoclonal and parallel monoclonal dissemination was considered together as monoclonal.

Using this alternate definition of dissemination, 33 tumors within the cohort were classified as polyclonal while the remaining 96 were classified as monoclonal using the definitions based on shared clonality. In order to perform the most accurate comparison, the tree inferred by the algorithm described in Chapter 3 was used as input and the hypothesis used by MACHINA was restricted to allow seeding from the primary tumor only (parallel single source seeding). Based on the input and hypothesis choice, MACHINA inferred the most parsimonious migration history and simultaneously, whether monoclonal or polyclonal seeding was more likely. In this case, 82 tumors were defined as monoclonal seeding from the primary tumor while 47 were defined as polyclonal seeding from the primary by MACHINA.
Figure 4-7 Comparison of dissemination patterns between TRACERx and MACHINA. The majority of tumors have consistent monoclonal or polyclonal calls across the two methods. Only a subset of tumors (14 tumors, in red) has differing calls.

For the majority of tumors (115 of 129, 89%) the dissemination calls were consistent across the two methods. The remaining 14 tumors were defined as monoclonal based on the adapted definitions of the method described above, whereas MACHINA defined these as polyclonal (Figure 4-7). This is consistent with the method described above being more conservative and under-calling heterogeneity. This discrepancy in inferring the dissemination patterns can be attributed to differing underlying assumptions of the two methods. MACHINA assumes that all metastatic sites have been sampled and therefore any clones present in all metastatic samples clonally must be shared with the primary – even if they are under the limit of detection. Therefore, MACHINA does not allow for metastasis-specific clones, but rather assumes that these clones are also present in the primary below the limit of detection, which alters the defined seeding events.

However, due to multi-region sampling of the primary tumor and leveraging this when calling mutations, as well as sufficient sequencing depth, the assumption for the method presented here was that any clones not found in the primary tumor were in fact absent from the primary tumor. Furthermore, due to clinical annotation it was known that not all
metastatic sites were sampled. Therefore, discrepancies between the two methods were to be expected.

Strikingly, when relaxing the hypothesis in MACHINA of allowing only seeding from the primary tumor, all 14 cases that differed across the two methods were inferred to have origins other than the primary tumor seeding some of the metastases. Therefore, all 14 cases (100%) were estimated to have metastasis-to-metastasis seeding. Conversely, of the remaining 115 tumors, only 7 (6%) showed evidence of metastasis-to-metastasis seeding as inferred by MACHINA.

### 4.2.6 Comparison of seeding clones to MACHINA

Additionally, the information about the seeding clones could also be extracted from the output provided by MACHINA and a comparison was performed to the clones defined as seeding based on the method described above. As MACHINA does not make the seeding clone information available directly, this information was extracted from the tree output, which classifies clusters based on which samples they originated from. Any change in classification from the primary to any metastasis was noted and the parental clone (i.e., the clone on the tree before the change in classification occurred) was classified as seeding. Similarly, when allowing seeding from sources other than the primary, any change in classification was defined as a seeding event and the parental clone extracted.

For the majority of cases (85/129 tumors, 66%; Figure 4-8), there was full overlap of the seeding clones between the two methods. However, in 21 cases (16%) none of the seeding clones overlapped (Figure 4-8). In 12 cases (9%) some of the defined seeding clones overlapped, but additional seeding clones were defined by MACHINA; in 3 cases (2%) the TRACERx method defined additional seeding clones; and finally, in 8 cases (6%) in addition to some overlapping seeding clones, both methods detected additional seeding clones (Figure 4-8).
Chapter 4 Results

Figure 4-8 Comparison of seeding clones between TRACERx and MACHINA.
The majority of cases have full overlap, while a significant subset shows no overlap. The right-hand plot shows the difference between the two methods in defining seeding clones split by overlap or discrepancies between the dissemination patterns as defined by the two methods.

Out of the 82 cases which were defined as monoclonal by both MACHINA and the TRACERx method (the latter also including some cases defined above as parallel monoclonal dissemination), in 67 tumors (82%) the seeding clones defined by both methods were identical (Figure 4-8). Of the remaining 15 tumors, no overlap of seeding clones was observed in 13 and 2 tumors had additional seeding clones defined by both methods. In these 15 tumors, all discrepancies resulted from MACHINA not allowing a metastasis unique clone, but rather classifying this as shared with the primary. Therefore, the seeding event was estimated to occur further along the phylogenetic tree, with the parental cluster being the one defined by TRACERx as the seeding clone. An example of this can be seen in Figure 4-9.
Figure 4-9 Example phylogenetic tree of tumor where TRACERx and MACHINA seeding clones differ. The left-hand side shows bar plots for each cluster in each tumor region. The mean CCF is noted within the bar plot. Additionally, the 95% confidence intervals are also shown. The right-hand side shows the phylogenetic tree inferred, as well as the clusters and number of mutations included. Each node of the tree is represented by a pie chart, which highlights which regions each cluster is present in, and a legend can be found in the top right corner. The seeding clone as defined by the TRACERx method and MACHINA are highlighted in blue and red, respectively.

This example, CRUK0010, has two primary regions (SU_T1.R1 and SU_T1.R2) and four metastasis samples (SU_FLN1, BR_LN1, BR_LN2, BR_LN3), one of which is a lymph node resected at surgery (SU_FLN1) and the remaining three are biopsies of lymph nodes at the point of disease recurrence. Using the method described above to determine the seeding clones, cluster 13 (light blue) and all subsequent clusters are defined as metastasis unique. Cluster 9 (light purple) is shared between all 4 metastasis samples and region 2 of the primary tumor. As this is the last shared cluster between the primary tumor and all metastases and is also clonal in all metastasis samples, this is defined as the only seeding clone by the TRACERx method.
Chapter 4 Results

For this example, MACHINA classifies cluster 13 as primary (all nodes highlighted in red originate in the primary tumor; Figure 4-10). Although tree resolution helps infer that cluster 13 is also present in the surgical lymph node (13\textsuperscript{p.lymphnode1}, highlighted in blue), as well as one of the recurrence biopsies (13\textsuperscript{relapse1}, highlighted in green), the seeding clone as defined by MACHINA is still cluster 13, inferred to originate in the primary tumor (Figure 4-10). For the remaining two recurrence biopsies (relapse2 and relapse 3), cluster 13 is not defined as present within the metastatic samples, but rather the seeding event for relapse2 is inferred as cluster 13 to cluster 11 (highlighted in the change of color from red to purple) and for relapse3 is inferred as cluster 13 to cluster 14 (highlighted in the change of color from red to orange; Figure 4-10).

For the 14 cases where the dissemination calls differed between MACHINA and the TRACERx method (see Section 4.2.5; Figure 4-7; Figure 4-8) the seeding clones were also compared. 8 tumors (57%) had no overlap of the seeding clones as defined by the two methods, 1 tumor (7%) had some overlapping seeding clones and also additional clones defined by both methods, while 5 tumors (36%) had overlapping seeding clones and MACHINA also defined additional clones as seeding compared to the TRACERx method. In these 14 cases the discrepancies could again be attributed to MACHINA not classifying the metastasis unique clone as such and therefore, the seeding clone was inferred as a child of the seeding clone as defined by the TRACERx method.

As mentioned in the previous section, all discrepancies between MACHINA and TRACERx in defining dissemination patterns could be explained due to a specific
assumption contained within MACHINA which does not allow for metastasis unique clusters. Indeed, in all 14 cases where the definitions differed were defined as metastasis-to-metastasis seeding when allowing for seeding from sources other than the primary. Therefore, the same approach was compared for seeding clones as well, to investigate whether a less restrictive hypothesis in MACHINA would lead to more overlap, or explain the discrepancies seen when comparing primary seeding only to the TRACERx method.

Interestingly, when comparing seeding from the primary tumor (primary seeding) and seeding from other sources (single source seeding) in the MACHINA analysis, the only changes in classifications occurred in cases that were inferred by MACHINA to have metastasis-to-metastasis seeding (9 cases with change of definition; 3 changed from both additional to MACHINA additional, 6 changed from no overlap to MACHINA additional; Figure 4-11). However, even when considering single source seeding, the discrepancies between MACHINA and TRACERx were still similar to those scenarios considering primary source seeding. No additional tumors had full overlap of seeding clones between the two methods (Figure 4-11). Overall, for 18 of the 21 cases with metastasis-to-metastasis seeding, MACHINA estimated additional seeding clones to the TRACERx method.
Finally, for the 33 cases that were defined as polyclonal by both MACHINA and the TRACERx method (for these cases at least one metastasis sample was defined as polyclonal at a sample level), the seeding clones fully overlapped across the two methods in 18 tumors (55%). In 5 tumors (15%) there was some overlap in the seeding clones defined, but both methods defined additional clones, in 7 tumors (21%) only MACHINA inferred additional seeding clones, while in 3 tumors (9%), the TRACERx method defined additional seeding clones not inferred by MACHINA. In these cases, the discrepancy was due to differences in the approach of defining the clone proportions while also defining the seeding clones. The input for MACHINA uses a bottom-up approach (estimating proportions from the leaf nodes back to the trunk of the phylogenetic tree; see Section 5.4.1) to define clone proportions for all clusters of the phylogenetic tree. This approach leads to some clusters being defined as absent, as the subsequent clusters have a cancer cell fraction equal to or greater than the one of the clusters in question. Therefore, this cluster might be defined as not present at the point of sampling. Conversely, in the TRACERx method, all putative seeding clusters are defined initially as clusters shared between the primary tumor and metastatic samples and subsequently, the clone proportions are calculated using only the cancer cell
fractions of those clusters. This results in a top-down approach, where clone proportions are calculated from the trunk to the leaf nodes of the phylogenetic tree. Due to variation and noise within the mutational data and the cancer cell fraction estimates, the two approaches can lead to differing definitions of presence or absence of clones which consequently can explain the discrepancies between MACHINA and TRACERx in these cases.

4.3 Quantifying selection in the seeding clone

Previous studies comparing primary and metastatic tumors have not identified many metastasis-unique cancer genes (Priestley et al., 2019; Nguyen et al., 2022). However, some differences between primary tumors and metastases were observed in various cancer types. Nguyen et al. found that metastases were significantly more chromosomally unstable and, in some cancer types, were more likely to show whole genome doubling compared to primary tumors (Nguyen et al., 2022). Additionally, the authors observed genomic features that were associated with metastases to specific organs (Nguyen et al., 2022). However, only single sample primary or metastasis tumors were considered for these analyses, with no paired primary-metastatic data.

By leveraging paired primary-metastatic data it was possible to define mutations that occur solely in the primary tumor and were not observed in the metastatic samples. Similarly, metastasis-unique mutations could be defined as mutations that were not detected in the primary tumor. All other events were classified as shared between the primary tumor and metastasis. Shared mutations could additionally be classified as truncal if they were shared across all regions of the primary tumor and metastatic samples. Finally, a subset of shared mutations could be defined as seeding (see Section 4.2.4) if they were present in the most recent common ancestor of the primary tumor and metastasis.
4.3.1 Cluster level: Seeding vs primary specific mutations

To investigate whether certain genomic events could be linked to metastatic potential, mutations from the seeding clone were compared to mutations from non-seeding clones. While all ancestral mutations, including the truncal mutations, would be present in the seeding clone, this analysis focused solely on the mutations that were unique to the seeding clone (this is referred to as the seeding cluster). The seeding cluster in early diverging tumors is necessarily the truncal cluster. Therefore, subsequent analyses were performed for all seeding clusters as well as only for subclonal seeding clusters.
To explore whether seeding clusters differed in size from their non-seeding counterparts, the mean cancer cell fraction across the primary tumor and proportion of primary tumor regions the seeding cluster was present in were considered. The mean cancer cell fraction across all regions of the primary tumor was significantly higher in the seeding clusters compared to primary unique clusters (median in seeding clusters: 0.475, median in primary clusters: 0.135; Wilcoxon rank sum test, $p = 2.28 \times 10^{-33}$; Figure 4-13A). This held true even when only considering subclonal clusters (median in seeding clusters: 0.313, median in primary clusters: 0.135; Wilcoxon rank sum test, $p = 1.27 \times 10^{-13}$; Figure 4-13B).

Furthermore, the proportion of primary tumor regions the seeding cluster was present in, was significantly higher than the proportion of primary tumor regions which harbored the primary specific clusters (median in seeding clusters: 0.667, median in primary clusters: 0.25; Wilcoxon rank sum test, $p = 2.33 \times 10^{-37}$; Figure 4-14A). When removing truncal clusters from the analysis the results remained significant (median in seeding clusters: 0.5, median in primary clusters: 0.25; Wilcoxon rank sum test, $p = 1.32 \times 10^{-17}$; Figure 4-14B). These results suggest that at the time of surgical resection, clones with metastatic potential are more likely to be dominant in the primary tumor. This could either reflect positive selection or neutral drift.

Figure 4-13 Comparison of mean CCF of the primary tumor for seeding and primary specific clusters.
The left panel shows the mean cancer cell fraction (CCF) in the primary across all clusters while the right panel is restricted to subclonal clusters only.

Furthermore, the proportion of primary tumor regions the seeding cluster was present in, was significantly higher than the proportion of primary tumor regions which harbored the primary specific clusters (median in seeding clusters: 0.667, median in primary clusters: 0.25; Wilcoxon rank sum test, $p = 2.33 \times 10^{-37}$; Figure 4-14A). When removing truncal clusters from the analysis the results remained significant (median in seeding clusters: 0.5, median in primary clusters: 0.25; Wilcoxon rank sum test, $p = 1.32 \times 10^{-17}$; Figure 4-14B). These results suggest that at the time of surgical resection, clones with metastatic potential are more likely to be dominant in the primary tumor. This could either reflect positive selection or neutral drift.
Positive selection can be caused by accumulation of somatic mutations in cancer genes which confer a fitness advantage and can lead to expansion of subclones. Conversely, in the absence of selection, i.e., during neutral evolution, subclones can grow at similar rates. In some cases, during neutral evolution, the size of subclones can vary due to random distribution of birth-death events (genetic drift) which is particularly pronounced at small population sizes (Ohta and Gillespie, 1996). Specifically, in the absence of selection, drift could lead to bigger clones, which just by chance are more likely to seed metastases, as they consist of more cells. Alternatively, bigger clones could represent “earlier” clones that have had more time to seed metastases.

Therefore, to evaluate whether the size of the seeding cluster reflects its fitness and therefore positive selection, the dndscv method (Martincorena et al., 2017) was applied to multiple gene lists. The dN/dS method assumes that if no selection is occurring, the rate of nonsynonymous and synonymous mutations should be equivalent. In this case, the number of nonsynonymous mutations per nonsynonymous site (dN) is proportional to the number of synonymous mutations per synonymous site (dS), or the ratio (dN/dS) is equivalent to 1. Furthermore, using this approach, a dN/dS value significantly >1 indicates positive selection. Firstly, a list of cancer genes described in (Martincorena et al., 2017) (Cancer genes Martincorena et al.; number of genes: 366), secondly, the list of genes described by the cancer gene census (Sondka et al., 2018) (version as of May
but excluding any that were associated with lung cancer (Cancer gene census no lung genes; number of genes: 469), thirdly, a curated list of lung cancer genes (Martincorena et al., 2017; Bailey et al., 2018) (Lung genes; number of genes: 111), and finally dndscv was run using all genes (All genes; number of genes: 20090) (see Section 2.2.4). The cohort was split by histological subtype and the dN/dS values were calculated for all mutations (global dN/dS). As the global dN/dS value includes all mutations (missense and nonsense mutations), it will reflect selection of both oncogenes and tumor suppressor genes.

For LUAD, significant positive selection of the cancer genes described by (Martincorena et al., 2017) could be observed across all types of mutations (dN/dS for seeding mutations: 1.38 [1.01-1.88], dN/dS for primary specific mutations: 1.24 [0.98-1.57], dN/dS for non-recurring mutations: 1.43 [1.14-1.8]; Figure 4-15). Similarly, for lung cancer specific genes, significant positive selection was observed in seeding and non-recurring mutations of LUAD tumors, although the estimate was slightly higher in seeding mutations (dN/dS for seeding mutations: 2.24 [1.28-3.91], dN/dS for primary specific mutations: 1.27 [0.86-1.87], dN/dS for non-recurring mutations: 2.08 [1.32-3.25]; Figure 4-15). No significant positive selection could be observed for non-lung cancer genes in LUAD tumors (dN/dS for seeding mutations: 1.1 [0.83-1.46], dN/dS for primary specific mutations: 1.07 [0.85-1.34], dN/dS for non-recurring mutations: 1.17 [0.96-1.43]; Figure 4-15). Conversely, for LUSC tumors significant positive selection was observed for non-lung cancer genes in seeding mutations only (dN/dS for seeding mutations: 1.82 [1.26-2.62], dN/dS for primary specific mutations: 1.17 [0.81-1.7], dN/dS for non-recurring mutations: 0.89 [0.67-1.19]; Figure 4-15). Similarly, significant positive selection of the cancer gene list described by (Martincorena et al., 2017) as well as lung cancer specific genes was observed only in seeding mutations in LUSC tumors (Martincorena et al. cancer genes: dN/dS for seeding mutations: 1.58 [1.11-2.25], dN/dS for primary specific mutations: 1.1 [0.74-1.64], dN/dS for non-recurring mutations: 0.87 [0.63-1.19]; lung cancer specific genes: dN/dS for seeding mutations: 2.06 [1.17-3.61], dN/dS for primary specific mutations: 1.5 [0.66-3.38], dN/dS for non-recurring mutations: 0.89 [0.54-1.49]; Figure 4-15).

Although not significant, a similar trend could be observed in lung cancer specific genes when only considering subclonal seeding mutations (Figure 4-16).
Figure 4-15 dN/dS comparison of mutations from seeding and primary specific clusters. Mutations from seeding and primary specific clusters from metastasizing tumors as well as mutations from non-recurring primary tumors were compared separately for LUAD and LUSC. The dN/dS values were calculated for all mutations (global dN/dS) using a general cancer gene list as described by Martincorena et al. (Martincorena et al., 2017), a list of cancer genes excluding lung cancer genes, a lung cancer specific gene list, and all genes.

Taken together, these results suggest that clones which seed metastasis are present at higher frequencies due to selection of cancer genes. This is particularly pronounced in LUSC tumors where significant positive selection is only observed in seeding mutations and not in primary specific mutations of metastasizing tumors nor in non-recurring tumors. Therefore, when analyzing unpaired data this signal of selection in LUSC tumors might not be observed.
Finally, to investigate whether specific genes were linked to positive selection in seeding and non-seeding clusters, the dN/dS analysis was repeated for mutations in seeding clusters, as well as primary specific and non-metastatic mutations combined (primary mutations) for all tumors in the cohort combined. The analysis was performed using the curated list of lung cancer specific genes and the gene level dN/dS estimate for each gene was considered for each of the two groups.

The odds ratio of each gene was calculated as the gene level dN/dS estimate of seeding mutations divided by the gene level dN/dS estimate of primary mutations. If the odds ratio was greater than 2, the gene was classified as seeding favored. Conversely, if the odds ratio was less than 0.5, the gene was classified as primary favored. Finally, if the odds ratio was between 0.5 and 2, inclusively, the gene was classified as seeding and primary. Only genes that had a q-value < 0.1 in at least one of the two groups were considered (Figure 4-17).

Additionally, to statistically compare the gene level dN/dS estimates between the two groups, a published approach as outlined by (Lawson et al., 2020) was used. The approach uses a one-sided likelihood-ratio test to compare dN/dS ratios between two datasets for a given gene $g$. 

---

Figure 4-16 dN/dS comparison of lung cancer genes for subclonal mutations only.

The plot shows dN/dS values separately for LUAD and LUSC tumors for a list of lung cancer genes. In this plot the mutations included are restricted only to subclonal mutations for seeding and primary specific mutations as well as non-recurring primary tumor mutations.
\[ H_0: \omega_{g,1} \leq \omega_{g,2} \]
\[ H_1: \text{unconstrained } \omega_{g,1} \text{ and } \omega_{g,2} \]

where \( \omega_{g,1} \) and \( \omega_{g,2} \) represent the dN/dS estimates for a given gene \( g \) in datasets 1 and 2, respectively. The approach corrects for differences in mutation densities related to coverage or mutational signatures, and also removes the effect of global differences in dN/dS ratios across the genes.

As the test is one-sided, it was performed twice, once comparing seeding versus primary mutations for genes significant in seeding mutations, and once comparing primary versus seeding mutations for genes significant in primary mutations. For both analyses, multiple testing correction was performed, and the significant genes (\( q < 0.05 \)) can be seen in Figure 4-17 highlighted in purple and green, respectively.

Figure 4-17 Gene level selection comparison between seeding and primary specific clusters. Seeding mutations are compared to primary-specific and non-metastasizing primary tumor mutations across all histologic subtypes. Genes are classified as seeding favored if the odds ratio (OR) of the dN/dS estimates of seeding vs non-seeding mutations >2, primary favored if the OR <0.5, and both primary and seeding
Significantly higher dN/dS ratios in seeding versus primary mutations could be observed in *NRAS*, *RBM5*, *TP53*, and *KRAS* (q-value: 0.0177, 0.0177, 1.96e-6, 0.177, respectively; Figure 4-17). For *NRAS*, 4 mutations were observed in seeding clusters while no mutations were observed in primary clusters. Therefore, the significant difference in dN/dS estimates highlights that although *NRAS* mutations in lung cancer are rare, when it is mutated, it seems to be associated with the clone that seeds metastasis.

The same method was applied for LUAD and LUSC separately and generated consistent results to the analysis which considered all tumor types together. In LUAD, *KRAS* and *TP53* were significantly enriched in seeding mutations even after multiple testing correction (q-value: 0.005, 0.002, respectively; Figure 4-18). No genes were significantly enriched in primary mutations in LUAD tumors after multiple testing correction. However, *DUSP22* and *SMAD4* were mutated predominantly in primary mutations compared to seeding mutations (*DUSP22*: 6 mutations in primary vs 0 mutations in seeding, *SMAD4*: 7 mutations in primary vs 0 mutations in seeding; q-value: 0.087, 0.087, respectively; p-value: 0.029, 0.023, respectively; Figure 4-18). This suggests that although *DUSP22* and *SMAD4* are positively selected in the primary tumor, they do not seem to confer a fitness advantage in the metastatic setting of the cohort presented here.
Figure 4-18 Gene level dN/dS analysis in LUAD.

Seeding mutations are compared to primary-specific and non-metastasizing primary tumor mutations across LUAD tumors. Genes are classified as seeding favored if the odds ratio (OR) of the dN/dS estimates of seeding vs non-seeding mutations >2, primary favored if the OR <0.5, and both primary and seeding otherwise. Genes highlighted in purple and green are significantly enriched in seeding and non-seeding mutations, respectively.

In LUSC tumors, TP53 was significantly enriched in seeding mutations (q-value: 1.68e-5) while B2M was significantly enriched in primary mutations and was not mutated in seeding mutations (5 mutations in primary vs 0 mutations in seeding; q-value: 0.0028; Figure 4-19). This suggests that HLA disruption via B2M mutations is not significantly selected at the metastatic transition in LUSC.
Chapter 4 Results

4.3.2 Paired analysis: Enriched vs depleted

Finally, to further evaluate individual cancer genes in the context of primary to metastatic transition, the multi-region paired primary-metastatic sampling in TRACERx was leveraged to perform a paired analysis of driver mutations.

Through the novel clustering and tree-building used to infer the evolutionary histories of mutations, it was possible to define each mutation cluster as clonal, subclonal or absent in each region (see Section 3.6.3). This classification could be summarized for a given
cluster across the primary tumor; if the cluster was defined as clonal in all regions of the primary tumor, the cluster was defined as clonal in the primary tumor overall. Conversely, if the cluster was absent from all regions of the primary tumor, the cluster was defined as absent from the primary tumor overall. Finally, if a cluster was classified as subclonal or absent in at least one region of the primary tumor, the overall classification was subclonal in the primary tumor. Analogously, the definitions could also be applied if multiple metastatic samples existed to summarize the metastasis level definitions. Consequently, each mutation of a given cluster could be defined as metastasis-overrepresented if the cluster was present at higher frequency in the metastasis compared to the matched primary tumor (i.e., absent in the primary tumor and subclonal or clonal in the metastasis, or subclonal in the primary tumor and clonal in the metastasis). Each mutation in a cluster could be defined as metastasis-underrepresented if it was present at lower frequency in the metastasis compared to the primary tumor (i.e., absent in metastasis and subclonal or clonal in the primary tumor, or subclonal in the metastasis and clonal in the primary tumor). Finally, each mutation in a cluster was defined as maintained if the cluster was present at similar frequency in both the primary tumor and metastasis (Figure 4-20).

![Figure 4-20 Schematic of definitions used in paired analysis.](image)

This classification was applied to all mutations across all paired primary-metastatic tumors. The background proportions of the classifications were calculated by considering all non-driver mutations, irrespective of the gene or tumor they occurred in. For this, all
LUAD (or LUSC) tumors and all non-driver mutations were considered, and the classifications added up across all these mutations, and finally divided by the total number of mutations considered for each cancer type.

To calculate the proportions for each cancer gene in each histological subtype (LUAD and LUSC), only driver mutations occurring in that cancer gene were considered and only cancer genes with more than 5 mutations (in LUAD or LUSC, respectively) were included in the analysis. For each gene and each histological subtype, the number of metastasis-overrepresented, -underrepresented and maintained mutations were calculated. These proportions and the proportion of such events in the background set of mutations are seen in Figure 4-21.

Finally, a multinomial test was performed to compare the classifications of each cancer gene included in the analysis to the background proportions of metastasis-overrepresented, -underrepresented and maintained mutations. Genes that were significant after multiple testing correction (FDR, q < 0.05) are shown in bold (Figure 4-21).

*Figure 4-21* Paired gene-level analysis. Analysis of mutations defined as metastasis-overrepresented, metastasis-underrepresented and maintained. Genes that were significantly different from the background proportion of classifications using a multinomial test are indicated in bold.
In LUAD tumors, KRAS and TP53 were significantly different from the background mutation classifications and more likely maintained (Multinomial test, FDR $p = 0.0036$, $p < 0.00005$, respectively; Figure 4-21). Similarly, in LUSC TP53 mutations were also significantly maintained (Multinomial test, FDR, $p < 0.00005$; Figure 4-21). This is consistent with the gene-level dN/dS results described above where TP53 had significantly higher dN/dS values in seeding versus non-seeding clusters (dN/dS 223.53 vs. 38.62, respectively; Figure 4-19). Only a single TP53 mutation was found subclonally in non-recurring primaries and no subclonal TP53 mutations were found in primary specific clusters of metastasizing tumors. This suggests that when TP53 is mutated subclonally, it is almost always associated with the clone that seeds metastasis.

Finally, in LUSC B2M mutations were significantly metastasis-underrepresented compared to background mutation classifications (Multinomial test, FDR $q = 0.0472$; Figure 4-21). This is again consistent with the gene-level dN/dS analysis where B2M is significantly selected in non-seeding clusters compared to seeding clusters (Figure 4-19). These results suggest that HLA disruption via B2M mutations is not significantly selected in the metastatic clone in LUSC.

### 4.4 Conclusions

The results of this chapter show that the predominant dissemination pattern within the TRACERx paired primary metastasis cohort is monoclonal dissemination. This is true both at the sample level, as well as the tumor level. Similar results have been observed in previous literature (Brastianos et al., 2015; Turajlic, Xu, Litchfield, Rowan, Chambers, et al., 2018; Hu et al., 2020; Lee et al., 2020). In other cancer types both monoclonal and polyclonal dissemination was observed (McPherson et al., 2016; Brown et al., 2017; Ullah et al., 2018; De Mattos-Arruda et al., 2019). However, in some studies, dissemination patterns were characterized using region-based phylogenetic trees and are therefore not comparable to the approach in defining dissemination patterns described here (Zhao et al., 2016; Brown et al., 2017; Naxerova et al., 2017; Reiter et al., 2020; Tang et al., 2021). Furthermore, by using region-based phylogenetic reconstruction methods, these studies were not able to define and characterize the seeding clone. This highlights the power of the approach to define dissemination patterns described here, which does not only consider similarity of mutations across the samples, but integrates clones and their phylogenetic relationships.
Chapter 4 Results

Consistent with previous literature (Reiter et al., 2020), primary LN metastases were enriched for polyclonal dissemination compared to recurrence/progression samples. This could indicate that primary LN metastases develop through a wider bottleneck, or that multiple seeding events are possible due to shorter anatomical distances between the primary tumor and primary LN compared to most recurrence/progression samples.

Furthermore, the results in this chapter show that the dissemination patterns could be influenced by under-sampling the number of metastatic sites. Indeed, the number of metastatic samples were significantly higher in paired tumor samples defined as polyclonal dissemination compared to monoclonal dissemination. This highlights that when only acquiring a single metastatic biopsy, heterogeneity might be under-called and clones seeding other metastases missed. Therefore, to explore the full range of metastatic heterogeneity and metastatic potential in the primary tumor, larger cohorts with extensive sampling of metastases, such as PEACE are needed. Similarly, other sampling methods such as ctDNA could aid in better characterizing the metastatic process (Abbosh et al., 2017; Parikh et al., 2019).

Finally, selection within the seeding clone was explored. Previous studies have compared primary and metastatic samples, however in an unpaired manner (Priestley et al., 2019; Nguyen et al., 2022). No clear differences were observed between primary tumors and metastases and almost no metastasis-unique drivers were identified. By leveraging the paired primary metastasis data, the metastatic seeding clone was defined as the last shared clone between the primary tumor and metastasis, therefore allowing for exploration of the processes within the primary tumor, including shared driver mutations, leading to metastasis. Subsequently, the seeding mutations were defined as those occurring solely in the seeding clone rather than the ancestral clones. Seeding mutations were compared to non-seeding mutations, which included mutations occurring solely in the primary tumor within the cohort as well as all subclonal mutations from non-metastasizing primary tumors. The results shown in this chapter show that seeding clones appeared to be fitter than non-seeding clones, including present in a larger proportion of primary tumor regions as well as at significantly higher cancer cell fractions. At the cohort level, significant positive selection of the seeding mutations was observed. This was particularly marked in LUSC tumors where no significant positive selection was observed in non-seeding mutations. Some differences were also observed at the gene
level between seeding and non-seeding mutations, including significantly higher selection in NRAS, RMB5, KRAS, and TP53 in the seeding compared to non-seeding mutations, and significantly higher selection of B2M in non-seeding compared to seeding mutations. B2M was also significantly metastasis-underrepresented in LUSC in the paired analysis further highlighting that HLA disruption via B2M mutations is likely not significantly selected in the metastatic clone. Similar results were observed in a previous study in colorectal cancer (Tikidzhieva et al., 2012). However, the overall number of mutations considered for the gene level and paired analyses were small. Therefore, a larger paired cohort would be needed to further explore patterns within the primary tumor leading to metastasis.

In this chapter I have mainly explored seeding patterns from the primary tumor and the seeding clone. However, it has previously been shown that not all metastases are seeded directly from the primary tumor. Therefore, in the next chapter I will further explore the metastatic process, including metastasis-to-metastasis seeding.
Chapter 5 Characterizing the metastatic process

5.1 Introduction

In the previous chapter I explored the process of metastatic dissemination from the primary tumor and characterized the metastatic potential of seeding clones within the primary tumor. In order to further characterize the metastatic process, it is also important to consider the timing at which the metastatic cells diverged from the primary tumor. As only a snapshot of primary tumor and metastatic evolution are captured at the time of sampling, it is not possible to determine when the metastatic cells disseminated from the primary tumor, but rather only the timing of divergence of the cell populations can be inferred.

Previous literature has characterized both early and late divergence of metastases (Turajlic and Swanton, 2016). Several studies in esophageal, pancreatic, and breast cancer have found predominantly late divergence of metastases (Yachida et al., 2010; Yates et al., 2017; Noorani et al., 2020). Conversely, studies in breast and colorectal cancer have found predominantly early divergence of metastases, however, these results could be biased by under-sampling of the primary tumor as well as using region-based phylogenies to infer the timing of divergence.

To further characterize the metastatic process, an extensively sampled metastasis cohort is introduced. PEACE (Posthumous Evaluation of Advanced Cancer Environment) allows for extensive sampling of metastatic sites and therefore a more complete picture of metastatic disease can be inferred. The data from PEACE can be used to explore intra- and inter-metastatic heterogeneity, as well as inter-site heterogeneity by leveraging genomic and transcriptomic data, as well as integrating detailed imaging and clinical annotations.

Finally, although dissemination patterns from the primary tumor were investigated in the previous chapter, several studies have shown that not all metastases are seeded directly from the primary (Gundem et al., 2015; Brown et al., 2017; Chen et al., 2022). Of particular interest are primary LN metastases and their role in seeding subsequent recurrences. Several studies have not found clear evidence that primary LN metastases
Chapter 5 Results

seed subsequent recurrences and distant metastases (Haffner et al., 2013; Naxerova et al., 2017; Ullah et al., 2018; Noorani et al., 2020; Tang et al., 2021).

In the following chapter, I will introduce timing of metastatic divergence and exploring these definitions in the TRACERx paired primary metastasis cohort. Additionally, I will introduce an extensively sampled metastasis cohort that can further help elucidate the metastatic process. Finally, I will explore metastasis-to-metastasis seeding in TRACERx and PEACE.

The work described in the following chapter is part of a manuscript currently under review. The TRACERx paired primary metastasis and the PEACE data were collected by members of the TRACERx and PEACE consortia, which I am also part of, and processed by me through a pipeline developed and adapted by members of the TRACERx consortium, including myself. The clinical annotations and imaging data for the PEACE cohort were collated by Sonya Hessey. Oriol Pich helped with the analysis of timing metastatic divergence using the platinum chemotherapy signatures. All other analyses described in the following section were conducted by me, if not otherwise stated.

5.2 Timing of metastatic divergence

Timing the divergence of the metastatic samples from the primary tumor could influence patient management. Metastases can diverge early while the tumor only consists of a small number of cells or late, when the tumor is larger. In the case of early divergence, it is possible that the metastases diverged before the tumor was detected and were already expanding when the primary tumor was resected. Identifying features of primary tumors that are likely to diverge early compared to late could be leveraged to better stratify patients for adjuvant therapy.

5.2.1 Defining timing of divergence

During tumor development, cells can disperse from the primary tumor to form metastases in various organs in the body. Metastases can occur once or multiple times during tumor evolution. Multiple studies have attempted to define the timing of metastatic divergence and dissemination but have often lacked sampling of the primary tumor (Klein et al., 2002). More recent studies have included both the primary tumor and metastases and
have found both early and late divergence across various cancer types including esophageal, pancreatic and colorectal (Yachida et al., 2010; Hu et al., 2019; Noorani et al., 2020).

Due to lack of longitudinal sampling of the primary tumor, timing divergence in absolute time is often not possible and the exact timepoint of dissemination cannot be pinpointed. In the absence of absolute timing of divergence, a few approaches can be used to approximate when the metastases likely diverged.

One approach is timing metastatic divergence relative to the last clonal sweep (Figure 5-1). If cells disseminate after the last clonal sweep this is referred to as late divergence. In this case, all mutations that are shared by all cells in the primary tumor are also shared by all cells in the metastases and the cells seeding the metastasis are subclonal within the primary tumor. Conversely, if cells disseminate before the last clonal sweep this is referred to as early divergence. In the case of early divergence, not all mutations defined as clonal within the primary are present within the metastatic sample. This means that the number of truncal mutations is lower than what would be estimated using the data from the primary tumor alone.

Other approaches to timing divergence include using mutational signatures (Alexandrov et al., 2013, 2020). Some mutations accumulate at a nearly constant rate throughout the life history of a tumor and are correlated with age in most cancer types (Alexandrov et al., 2020) therefore, these mutations can be used to infer the timing of metastatic divergence. However, when using whole exome sequencing, not enough such mutations are detected and therefore it is difficult to infer timing in such cases. Additionally, it is possible to use mutations attributed to smoking to time the divergence of metastases in ex-smokers. However, this again is limited to a smaller cohort of cases and is only possible in a subset of these where smoking related mutations can be detected in metastasis specific mutations. Similarly, certain chemotherapy agents have been shown to lead to higher mutation burden and have been associated with specific mutational processes (Johnson et al., 2014; Pich et al., 2019; Alexandrov et al., 2020). Although the power to detect such mutations is again limited using whole exome sequencing, in cases where a platinum chemotherapy signature can be detected, it is possible to leverage this to time metastatic divergence.
5.2.2 Characterizing early/late diverging metastasis

When classifying the timing of divergence based on mutations alone, some bias can be introduced, for example due to subsequent copy number loss in some regions of the primary tumor. If a stringent threshold for calling mutations is used, only mutations with very high confidence would be called. This might result in misclassification of some mutations as absent in the metastatic samples. By lowering the threshold to define mutations as present in a given sample, additional mutations could be salvaged.

By relying on the clusters defined by the updated clustering process, the presence/absence calls have been corrected for both purity and ploidy and subclonal
copy number losses are taken into account. Therefore, instead of considering presence or absence of individual mutations in any given sample, presence of groups of mutations is considered.

For this approach, all the mutation clusters present on the phylogenetic tree were considered. Each cluster was defined as clonal, subclonal or absent in both the primary tumor and metastases. Only clusters defined as clonal within the primary tumor were subsequently considered. If all clusters defined as clonal within the primary tumor were also clonal in the metastases, then all clonal mutations were shared, and the tumor was defined as late divergence. Conversely, if not all the clusters defined as clonal in the primary tumor were clonal in the metastatic samples, this was defined as early divergence. In this case the proportion of shared mutations could be calculated to give an indication of how many primary-clonal mutations were missed.
No significant difference in proportions of early and late divergence for LUAD and LUSC were observed, however a slightly higher proportion of early divergence cases were observed in LUSC compared to LUAD.

Overall, in the TRACERx paired primary metastasis cohort, 73.6% (95/129) of the cases showed evidence of late divergence, while only 26.4% (34/129) were defined as early divergence (Figure 5-2). When considering the main histologic subgroups separately, a slight difference was observed in the proportion of early to late divergence between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) with a larger proportion of early versus late divergence observed in LUSC (LUSC: 16 early vs 24 late; LUAD 14 early vs 52 late; Fisher’s exact test, p = 0.047; Figure 5-3). However, when also considering other histologic subtypes, the difference in proportion was no longer significant (Fisher’s exact test, p = 0.063). Finally, no difference in timing patterns was observed when considering primary LN and subsequent recurrences (Fisher’s exact test, p = 0.75; Figure 5-4).
As in early divergence cases not all primary clonal mutations are shared with metastases, treatments targeting primary-clonal mutations might fail in such cases. Therefore, it is important to further understand which mutations are shared between the primary tumor and metastasis as these represent putative targetable mutations. Importantly, even in early divergence cases, the majority of primary-clonal mutations was found to be shared with metastases (Figure 5-5). This suggests that although metastatic divergence might occur prior to the last clonal sweep, it still occurs relatively late in the evolution of the tumor. Furthermore, when considering primary-clonal mutations occurring pre-whole genome doubling (WGD) these were less likely to be absent in metastases compared to post-WGD mutations as well as mutations occurring in non-WGD tumors (median percentage of absent pre-WGD mutations: 1.67%, IQR: 0-6.49%; median percentage of absent post-WGD or non-WGD mutations: 8.72%, IQR: 2.71-30.48%; Wilcoxon signed-rank test, p = 0.0056; Figure 5-5).
In early divergence cases, the median number of pre-GD mutations defined as absent/not clonal in the metastases is 1.67% IQR [0-6.49%]. Post-GD mutations or mutations in non-GD tumors were more likely to be absent in the metastasis (median 8.72% IQR [2.71-30.48%]; Wilcoxon signed-rank test, p=0.0056).

5.2.3 Sampling bias due to number of primary tumors sequenced

The methods described above to determine the divergence of metastasis relative to the last clonal sweep are dependent on the correct classification of clonal mutations within the primary tumor. The metastases are classified as late diverging only if all mutations defined as clonal in the primary are also clonal in the metastasis. However, it is possible that some of the mutations that were defined as clonal within the primary were only defined as such due to clonal illusion driven by under-sampling of the primary tumor (de Bruin et al., 2014). For example, if only two regions of a primary tumor are sampled, all mutations occurring in both regions would be defined as clonal within the primary tumor. If subsequently a third region was sampled, some mutations which were found in both regions 1 and 2 could be absent from region 3, and therefore these would then be classified as subclonal in the primary tumor.

For this reason, a down sampling analysis was performed to quantify the bias introduced by tumors with lower number of regions sampled. This was done for all cases that were classified as late diverging to see how many regions were required to accurately classify these cases as late diverging. When down sampling to a single region, all clusters that were defined as clonal in a given primary region were considered and, equivalent to above, if all clusters were also found to be clonal in the metastatic samples, the case was defined as late divergence. If not all clusters defined as clonal in the single primary region were clonal in the metastatic samples, the case was defined as early. This was repeated for all regions of the primary tumor, averaging the proportion of shared mutations across all regions to determine the final proportion of shared mutations and define the timing. Therefore, for a single region the case was only defined as late if all primary tumor regions individually were defined as late. The same approach was repeated for more regions. Here, only the clusters found to be clonal across all regions were then compared to the metastatic samples. Tumors with n regions were only considered for down sampling up to n-1 regions. For example, if a tumor only had three primary tumor regions sampled, it was only considered in the down sampling analysis for one and two regions.
Strikingly, when only sampling one region of the primary tumor, 78% of late divergence cases were misclassified as early divergence (Figure 5-6). This proportion shifts significantly when sampling two regions of the primary tumor, with over 50% of cases properly classified as late. However, in some cases even with three or four regions sampled, the divergence is still misclassified as early. This highlights the importance of multi-region data when characterizing timing of divergence relative to the last clonal sweep.

In a study of colorectal cancer, the authors inferred predominantly early divergence with 18 out of 23 metastases from 17 patients defined as early divergence. Strikingly, four out of five metastases from three out of four tumors estimated as late divergence had multi-region sequencing of the primary tumor compared to only six of 17 tumors estimated as early divergence (Hu et al., 2019).

Figure 5-6 Under-sampling the primary tumor can lead to incorrect classification of timing of divergence. For cases defined as late divergence, down sampling of the primary tumor regions was performed. The top panel depicts the proportion of cases classified as early versus late when only sampling a subset of regions. The bottom panel shows the proportion of shared mutations for each of the cases.
5.2.4 Timing divergence using mutational signatures

As mentioned above, timing metastatic divergence in absolute time can be challenging, especially using whole exome sequencing data. In the paired primary metastasis cohort, sequencing data of the recurrence or progression was available for 27 patients who had received adjuvant platinum chemotherapy. The platinum mutational signature could only be detected in metastasis unique mutations of 7 of the 27 patients (26%) with 19 of 20 patients where no signature was detected having low mutation burden. In one of the 7 patients where a platinum mutational signature could be detected, these mutations could be used to time metastatic divergence. This patient, CRUK0590 presented with two separate brain metastases at disease relapse. One brain metastasis in the posterior fossa and cerebellum had two regions sampled, while the other brain metastasis had a single region sampled. A platinum chemotherapy signature could be confidently detected in the second brain metastasis, but in neither region of the other brain metastasis. Furthermore, the mutations shared between both brain metastases also showed no evidence of a platinum chemotherapy signature. Therefore, these metastases must have diverged prior to treatment, which occurred approximately 6-8 months prior, with one metastasis subsequently acquiring unique mutations that could be attributed to the platinum chemotherapy treatment.
Figure 5-7 Example of timing divergence using mutational signatures. An active platinum chemotherapy signature was detected in one brain metastasis sample, while in the two regions of the other brain metastasis (posterior fossa and cerebellum) no platinum chemotherapy signature was observed. This suggests that the two brain metastases diverged prior to platinum therapy being delivered, approximately 6 to 8 months prior to the recurrence.

5.3 Overview of PEACE cohort

To better explore the metastatic process, a national prospective clinical study was set up integrating multi-omics as well as the immune microenvironment to study cancer evolution. Some of the patients recruited in TRACERx were co-recruited into PEACE allowing for genomic analysis of at least two timepoints as well as multiple sites of disease. To date, samples from 22 TRACERx patients have been sequenced as part of PEACE. An overview of the samples and timepoints obtained from the patients can be found in Figure 5-8. At tissue harvest, metastases were obtained from a total of 18 different tissue types. The largest group of metastatic samples sequenced at tissue
harvest were obtained from lung metastases (123/348 samples; 35.3%), with metastases from lymph nodes making up the second largest group (49/348 samples; 14.1%; Figure 5-9). Other sample types with multiple metastases sequenced included liver (34/348; 9.8%), peritoneum (29/348; 8.3%), as well as soft tissue and adrenal (both with 18/348 samples; 5.2%; Figure 5-9).

To accommodate the large number of samples as well as the multiple timepoints present within the PEACE cohort, the pipeline described in (Jamal-Hanjani et al., 2017) was adapted to create a PEACE specific pipeline. Firstly, each set of samples sequenced separately, was sequenced with a paired germline. Therefore, samples obtained from each timepoint needed to be processed independently. In some cases, additional samples from tissue harvest were sequenced at a later date, meaning additional germline samples were present. The alignment, variant calling, SNV calling, and copy number inference using ASCAT (Van Loo et al., 2010) were performed separately for each paired germline.

After manual quality control of purity and ploidy solutions, 34 samples were failed due to low tumor purity (< 5%) and low mutation burden and excluded from further analysis. Additionally, 5 samples had sufficient mutation burden, but low tumor purity and were
therefore not included in the clustering and tree building. For one patient (PEA052) all samples collected at tissue harvest failed manual quality control due to high level of necrosis and therefore the patient was excluded from the cohort. Therefore, the final number of patients included was 21 and the number of tissue harvest samples included for all subsequent analyses was 309.

![Figure 5-9 Overview of tissue types sampled within PEACE.](image)

After these initial processing steps, all data were combined and the multi-region sequencing data was leveraged. All variants detected across all samples were pooled together, and coverage, variant and reference reads were re-extracted using bam-readcount (https://github.com/genome/bam-readcount) across all samples and timepoints. Subsequently, copy number correction and phasing of alleles was performed across all samples using only overlapping SNVs found across all samples and germlines.

A clustering step was performed to detect whether samples shared a clonal origin. If at least 10 ubiquitous mutations existed across all regions, the samples were defined as a single related tumor, with a shared clonal origin. If, however, less than 10 ubiquitous mutations were found, VAFs across all regions were clustered using the R function hclust. The clustering tree was split into two groups determining the samples associated with the two distinct tumors. In three PEACE cases, multiple tumors were detected. For
one case, PEA110 there was a second primary detected at surgery from which a single sample was obtained and processed. This second primary was a pleomorphic giant cell carcinoma which did not go on to metastasize, while the primary adenocarcinoma seeded all sites of disease. The single-sample second primary was excluded from further analysis. Similarly, for PEA143 and PEA102, a second genomically unrelated tumor was detected. For both cases this consisted of a single metastatic sample: oesophagus2 and lung1, respectively. Both of these samples were not included for further analysis. Additionally, to reduce the heterogeneity in these cases, two regions were excluded, as they shared a lower proportion of ubiquitous mutations with the remaining samples (PEA102_TH_BO03 and PEA110_TH_LU27).

Finally, to account for the large number of samples and mutations, the clustering approach as previously described in (Jamal-Hanjani et al., 2017) was updated as described in Chapter 3.
Figure 5-10 Genomic overview of PEACE cohort.

The top panel shows the number of primary-specific, shared and metastasis-specific mutations for each patient within the cohort, while the second panel shows the proportions of mutations across the three categories. Patients are ordered based on the proportion of shared mutations. The third and fourth panel
A genomic overview of the cohort can be found in Figure 5-10. The median mutation burden was 1305 (inter-quartile range [IQR]: 942 – 1576). The number of mutations shared between the primary tumor and metastases varied across cases, with only 69 mutations shared in PEA289 yet 1572 mutations shared in PEA013 (median number of shared mutations: 443; IQR: 209 – 921; Figure 5-10). Similarly, the number of metastasis unique mutations varied substantially across cases (median number of metastasis unique mutations: 611; IQR: 225 – 851; Figure 5-10). Conversely, the number of primary unique mutations varied much less, with a median of 144 mutations only found in the primary tumor (IQR: 91 – 212; Figure 5-10).

Driver mutations were detected in every case, with a median of 8 mutations per case defined as drivers (IQR: 5 – 12; Figure 5-10). A median of 4 driver mutations per case were shared between the primary tumor and metastases (IQR: 2 – 7; Figure 5-10). A median of 2 driver mutations per case were only found in metastases (IQR: 1 – 4) while a median of one driver was found to be unique to the primary tumor (IQR: 0 – 1; Figure 5-10). The most frequently mutated gene was TP53, with 14 mutations shared and only one found to be metastasis unique. The second most frequently mutated gene was ARID1A (4 shared mutations, 2 metastasis unique mutations). In three cases, multiple driver mutations occurred in the same gene: in PEA037, both a primary unique and a metastasis unique driver mutation was observed in PTEN indicating parallel evolution in both the primary and metastasis; in PEA008 a shared driver mutation as well as a metastasis unique driver mutation in EP300 were found, indicative of a second hit in the metastatic samples; in PEA141 two B2M driver mutations were observed, one shared and one primary unique. This finding suggests that a double hit in the primary tumor might confer a fitness advantage, while in the metastatic setting disruption of the antigen presentation mechanism through B2M mutation might not be significantly selected. Similar results were observed in the TRACERx paired primary metastasis cohort (se Section 4.3).
Figure 5-11 Overview of all PEACE trees.
Trees for all patients within the cohort are shown. Clusters on the tree are colored by the tissue type they appear in. If clusters are shared across multiple tissue types, they are shown in light blue.

An overview of all trees within the PEACE cohort can be found in Figure 5-11. Nodes are colored based on the tissue type they are associated with, which can give an indication which tissue types are more similar or different. Similarly, the nodes can be colored by the timepoint they occur in, an example for PEA013 can be found in Figure 5-12. The tree on the left shows all clusters that were present in the primary tumor, the tree in the middle shows all clusters present in the recurrence biopsy from the left adrenal, while the tree on the right shows all clusters present at tissue harvest, colored by the tissue they were present in. This gives an indication of sampling bias: in the middle tree, highlighting clusters found in the left adrenal relapse, only the left branch of the phylogenetic tree is highlighted as present, however in the right tree showing tissue harvest samples, the right branch is once again highlighted and present in the liver samples (nodes colored in brown). This means that either the liver metastases were seeded early by the primary tumor and were not sampled at recurrence, or that there
was some residual disease that seeded the liver at a later timepoint, after the initial recurrence. For some samples, like the liver metastases, the origin can be inferred, as the clusters present in the liver (colored in brown) are most closely related to the primary tumor (colored in dark blue) and no other sites of disease share subclonal mutations with the liver. However, in most samples inferring their origin is more difficult. For example, the adrenal metastases (colored in pink) could have originated from the primary tumor, or from the recurrence.

![Figure 5-12 Timepoint trees of PEA013 with nodes colored that are present at primary surgery, recurrence, and tissue harvest.](image)

Analyzing the timing of divergence within the PEACE cohort, in the manner described above, yielded broadly consistent results to the TRACERx paired primary metastasis cohort with 33% (7/21 cases) defined as early and 67% (14/21 cases) defined as late divergence (Figure 5-13). As extensive sampling of metastases has been performed, this highlights that the sampling bias in determining the timing of divergence is predominantly based on the number of primary regions sampled and that sampling of additional metastases does not change the patterns of early and late divergence observed within the TRACERx cohort.
The majority of cases within the PEACE cohort are classified as late divergence, with a smaller subset defined as early divergence.

To further explore the timing of individual metastases, imaging data was included that give an indication of when a metastasis was initially clinically detected. While this is an approximation of the timing, the metastasis might have been present long before it was visible on a scan. Additionally, inconsistent durations between scans in some cases could mean that the metastasis was at a clinically detectable size, but only picked up on scans at a later date.

Furthermore, to infer the origin of metastatic sites, a published tool MACHINA (El-Kebir, Satas and Raphael, 2018) was applied. This approach can also help leverage timing of metastatic dissemination, as if one metastatic site seeds another then knowing the timing of either metastasis limits the emergence of the other.

### 5.4 Inferring migration histories

The methods described to determine the dissemination patterns as well as the timing of divergence only take into account divergence and dissemination from the primary tumor. However, a metastasis can also be seeded from an existing metastasis (metastasis-to-metastasis seeding) or even from multiple sources, including the primary tumor and other
sites of disease. To further explore the seeding source of each metastasis, a published tool MACHINA was implemented that infers migration histories from clone trees (El-Kebir, Satas and Raphael, 2018).

MACHINA can infer both the clone tree and migration patterns together. For this, estimates of the cancer cell fraction with lower and upper bounds needs to be provided as input. However, for the majority of the cohort this approach failed likely due to complex tree structures as well as wide confidence intervals of some clusters. Alternatively, it is possible to provide a clone tree as input and only infer the migration histories using MACHINA. When providing a clone tree, MACHINA can additionally perform tree resolution which will partially resolve polytomies (only those necessary to infer the migration histories, avoiding any redundancies) within the clone tree.

Furthermore, MACHINA can run under multiple hypotheses. The most restrictive hypothesis only allows seeding from the primary tumor. This means that no metastasis-specific clones can exist as all metastases are independently seeded from the primary tumor. The second possible hypothesis still only allows a single seeding source, but this is not limited to the primary tumor. Under this hypothesis for example, metastasis A can be seeded by the primary tumor while metastasis B could be seeded by metastasis A. The third hypothesis allows for seeding from multiple sources which means metastasis B can be seeded by both the primary tumor and metastasis A. Finally, the unrestricted hypothesis allows for reseeding which means that metastasis A can seed metastasis B which can also seed metastasis A. In all four hypotheses MACHINA directly infers the seeding clonality as either monoclonal, if a single migration event occurs to each metastasis, or polyclonal, if at least one metastasis was seeded by multiple migrations from the same source.

5.4.1 Inferring clone proportions

The clone tree, which is used as input for MACHINA, is characterized by a set of leaf nodes that are equivalent to the clones present in each region at the point of sampling. This is in contrast to the phylogenetic tree obtained from the tree building method which only highlights the cancer cell fractions of each cluster and a likely relationship between all clusters of a given tumor, irrespective of whether the cluster is actually observed or ancestral. However, by using the phylogenetic tree from the tree building, it is possible to infer the clone proportions of each cluster and subsequently infer the clone tree.
Given a tree topology and cluster CCF estimates, the clone proportions can be calculated as follows:

- For the leaf nodes of the tree, the CCF estimate is equivalent to the clone proportion as these are terminal clones.
- Using a “bottom-up” approach each parental cluster’s clone proportion equals the sum of the clone proportions of the children subtracted from the CCF of the parental cluster.

For example, if a cluster (CCF = 50%) has a single descendant (clone proportion = 30%), then the parental cluster’s clone proportion is 50 – 30 = 20%. If the parental cluster (CCF = 50%) has two descendants (clone proportion = 30% and 20% respectively) then the parental cluster is not present in the tumor, but rather an ancestral population that has fully been replaced by the two descendant clones.

- This approach is repeated through all tree levels ending with the truncal cluster.

Due to uncertainty when inferring the clusters as well as the CCF buffer used in tree building, the CCF point estimate of each cluster may not be fully accurate leading to a bias when inferring clone proportions. One approach to improve this, is by setting the CCF of all clusters defined as clonal in any given region to 100% as these clusters are indistinguishable from the truncal cluster. This correction reduces the number of clones inferred for a given tree as it removes any small clones due to minor differences in CCF.

For example, if two clusters, cluster1 and cluster2, have a CCF of 97% and 90% respectively then the clone proportion of cluster1 would be 7%. If both clusters were indistinguishable from the truncal cluster, then both CCFs would be corrected to 100% and therefore cluster1 would not be present within the sample.

Applying the same approach to all subclonal clusters is more difficult, as testing the difference of cluster CCF distributions is not a reflexive relation. This means that cluster 4 might be significantly different from the truncal cluster 1, while clusters 2 and 3 are not significantly different from the truncal cluster 1. However, cluster 4 is not significantly different from cluster 3, meaning it is both clonal and subclonal depending on the relationship tested.
For this reason, a bootstrapping approach was also implemented to infer clone proportions alongside an underlying distribution. This allows for testing whether a clone is present or absent in a given sample as well as estimating the maximum likelihood of the joint clone proportions in a given region. For this approach, \( n \) mutation CCFs are sampled with replacement from a given cluster with \( n \) mutations from a given region. Subsequently, the mean CCF of the sampled mutations is calculated. This is repeated for all clusters in a given region and the mean CCFs are used to infer the clone proportions as described above. This approach is applied across all regions to estimate all clone proportions. Finally, this is repeated \( N \) times for each region and all clusters in that region to estimate the distribution of clone proportions. To then test presence/absence of a clone, the number of occurrences where the clone proportion is 0 are counted and if the proportion of occurrences out of the \( N \) repeats falls below a p-value threshold (for example, 0.01), then this clone is defined as present.

### 5.4.2 Implementing and running MACHINA

The clone tree needed as input for MACHINA can be inferred from the calculated clone proportions or presence/absence definitions from the bootstrapping approach. For this, all clusters present as clones in a given region are added as additional nodes to the original phylogenetic tree inferred by the tree building method. To avoid double counting clones only present in a single region, leaf nodes only present in one sample are removed and only the extant clone inferred is kept.

To run MACHINA, additionally, the mapping of extant clones to regions is also provided. MACHINA is run using the unconstrained hypothesis (allowing for reseeding) as in this scenario all other hypotheses are also run. This means that seeding patterns are inferred, allowing only for single source seeding from the primary, from other single sources, multiple sources as well as reseeding and the most likely solution for each scenario is provided. These results can then be summarized and the region of origin as well as the seeding clone extracted from the output.

### 5.4.3 Quantifying occurrences of metastasis-to-metastasis seeding in TRACERx

When running MACHINA under less restrictive hypothesis, allowing seeding from a single source that is not necessarily the primary, or even allowing multi-source seeding or reseeding, it is possible to start exploring whether one metastasis is likely seeded by
another metastasis. To perform this analysis, multiple metastasis samples of a given tumor are necessary. Therefore, this analysis was restricted to 51 cases with multiple metastasis samples from the paired primary metastatic cohort. Of these, 26 patients had multiple lymph nodes resected at surgery, 22 patients had both lymph nodes resected at surgery as well as subsequent recurrence or progression samples, and 3 patients had multiple recurrence or progression samples only.

![Figure 5-14 Overview of metastasis-to-metastasis seeding in TRACERx.](image)

The left plot shows the number of patients with seeding solely from the primary tumor versus patients where some metastasis-to-metastasis seeding is observed. The right plot separates patients further into those where only primary LN metastases were sampled, those with both primary LN and subsequent recurrence/progression samples, and those with only recurrence/progression samples and highlighting the origins of metastases within all cases.

Overall, metastasis-to-metastasis seeding was potentially present in 21 of 51 cases (41%; Figure 5-14). Of the 26 patients with multiple lymph nodes sampled at surgery, lymph node to lymph node seeding was observed in 13 cases (50%), while in the remaining 13 cases all lymph nodes were seeded independently from the primary tumor (Figure 5-14). Of the 13 cases with lymph node to lymph node seeding, 9 patients had a subsequent recurrence of their disease. For all these patients the recurrence was not sampled, therefore is not possible to discern whether the potential of the lymph nodes to seed other lymph nodes would have also allowed subsequent seeding of the recurrence samples. However, of the 13 patients without lymph node to lymph node seeding, 8 developed a subsequent recurrence, which could mean that lymph node to lymph node
seeding at the time of diagnosis is not likely to be a good indicator of subsequent disease progression.

This is further highlighted by the fact that of the 22 cases with both lymph nodes resected at surgery and subsequent recurrence or progression samples, metastasis-to-metastasis seeding could only be observed in 7 cases (32%) only 4 (18% of the 22 cases) of which could be attributed to lymph node to recurrence seeding (Figure 5-14). In the remaining 3 cases where metastasis-to-metastasis seeding could be observed, in one case, CRUK0029 (Figure 5-15), the primary tumor seeded two lymph nodes (p.lymphnode5 and p.lymphnode8) as well as the metastasis at relapse (relapse1); one of the lymph nodes (p.lymphnode8) seeded an additional lymph node sampled at surgery (p.lymphnode7).

In another case, CRUK0242 (Figure 5-16), the primary tumor seeded the metastasis at relapse (relapse1), which in turn seeded the lymph node that was previously resected at surgery (p.lymphnode1).
In the final case where metastasis-to-metastasis seeding could be observed, CRUK0484 (Figure 5-17), the primary tumor independently seeded two lymph node metastasis sampled at surgery from distinct subclones within the primary tumor, as well as a progression in the scapula (day 433) and a subsequent brain progression sample (day 582) from two additional distinct subclones. Finally, the scapula metastasis seeded a rib metastasis which was the initial recurrence metastasis detected (day 147).

Of the three cases with only recurrence or progression samples, only one showed evidence of metastasis-to-metastasis seeding. In this case, CRUK0090 (Figure 5-18), the primary tumor seeded a lymph node at progression (progression1) which seeded a second lymph node sampled at that time point (progression2), as well as the previously detected lymph node at recurrence (relapse1). Interestingly, all these samples were derived from a cluster of lymph nodes at station 4R.
5.4.4 Quantifying occurrences of metastasis-to-metastasis seeding in PEACE

The same approach as described above was applied to the PEACE cohort of 21 patients. Overall, for 5 patients MACHINA inferred seeding from the primary tumor alone, meaning no metastasis-to-metastasis seeding was observed. In the remaining cases, some form of metastasis-to-metastasis seeding was inferred. In four cases multi source seeding was observed, however in one of these cases this was only detected when running the unrestricted method rather than just allowing for multi-source seeding. Finally, when running the unrestricted model, in one case (PEA141) reseeding was inferred (Figure 5-19).
In this example, the primary tumor (located in the left lower lobe) seeded three separate lymph nodes resected at tissue harvest (lymphnode3, 4, 5). Lymphnode 3 and 4 were both sampled from the right hilum, while lymphnode 5 was a subcarinal lymph node. Additionally, lymphnode3 was also seeded by lymphnode4 (highlighting multi source seeding of this lymph node). Lymphnode4 also seeded the initially detected recurrence (relapse1). Lymphnode1 (sampled from the left hilum) which had reseeding with lymphnode3 (sampled from the right hilum) then went on to seed both lymphnode2 (sampled from the left hilum) and lymphnode5 (subcarinal lymph node), which was also seeded directly from the primary tumor, again showing patterns of multi-source seeding.

Overall, the majority of metastatic tumors are seeded directly from the primary tumor (115/226; 51%; Figure 5-20). Metastatic sites seeding large number of additional metastases include lymph nodes (44/226; 20%), lung (34/226; 15%), and liver (11/226; 5%; Figure 5-20).

Figure 5-20 Overview of different tissue types seeding further metastases.

The majority of metastases are seeded directly from the primary tumor, however a significant subset are seeded by other sources. Metastases seeding the largest number of subsequent metastases are lymph nodes and lung metastases.
Lymph node metastases can go on to seed various tissue types, including further lymph node metastases, lung, liver, brain, and peritoneal metastases (Figure 5-21). Similarly, lung metastases can give rise to lung, lymph node, liver, bone, and peritoneal metastases (Figure 5-21). While lymph node and lung metastases give rise to a high proportion of different tissue types, liver metastases generally seed other liver metastases compared to other sites of disease (Figure 5-21). When metastases from the brain or soft tissue seed further metastatic sites, they only seed additional metastases in the same organ (Figure 5-21). Although the number of samples and migrations is not large enough to discern significant recurrent patterns, the data gives an indication of possible routes of metastases.
Figure 5-21 Alluvial plot highlighting source and targets of different migration patterns.
5.4.5 Timing metastatic migrations

For one patient, PEA013 extensive clinical and imaging annotation was obtained. The patient was a 60 year old male that presented with stage 3a squamous cell carcinoma of the right upper lobe. During the lobectomy, 8 primary tumor samples were obtained and sequenced. On a CT scan 50 days after surgery, a suspicious mass in the left adrenal as well as retroperitoneal lymph nodes were detected. Nine days later this was confirmed using PET-CT, where additional metastases were found in the T8 and L4 vertebra as well as in the sacrum as well as the left frontal brain (Figure 5-22). The brain and vertebral metastases were further confirmed with an MRI 11 days later. A biopsy of the left adrenal metastasis was obtained and sequenced. The patient then received multiple lines of chemotherapy, and radiotherapy localized to the 4th rib and T8 bone metastases (Figure 5-22). On a subsequent MRI approximately 40 days later, the left frontal brain metastasis had progressed further, and a new right occipital lobe metastasis was detected. On the final CT scan 139 days after surgery, a new right adrenal metastasis was found, and the left adrenal metastasis had increased significantly in size (Figure 5-22). The patient died 251 days after surgery and during tissue harvest, 26 metastatic samples were obtained.

Figure 5-22 Clinical and imaging timeline of PEA013.

MACHINA predicted independent seeding events from the primary leading to the two brain metastases in the left frontal and right occipital lobe (brain1 and brain2, respectively), a 4th rib metastasis (bone1), an adrenal adjacent lymph node (lymphnode1), as well as the left adrenal metastasis (adrenal1) (Figure 5-23). The left
adrenal metastases sampled at tissue harvest (adrenal1) subsequently seeded the
initially sampled recurrence biopsy of the left adrenal metastasis (relapse1), the
metastasis in the L4 vertebra (bone2) two retroperitoneal lymph nodes (lymphnode2 and
lymphnode3), as well as the right adrenal metastasis (adrenal2) (Figure 5-23).
Interestingly, the adrenal adjacent lymph node and the left adrenal metastasis were
independently seeded by the primary tumor in two separate events.

When tying the migration patterns back to the imaging data, it is possible to infer the
relative timing of presence of certain metastases. For example, the L4 vertebra (bone2)
which was seeded by the left adrenal metastasis (adrenal1; Figure 5-23) was detected
on scans 50 days after the initial surgery, at the same time as the left adrenal metastasis
was also discovered (Figure 5-22). This means that the left adrenal metastasis must
have already existed significantly prior to these scans and had sufficient metastatic potential to seed the bone metastasis.

Furthermore, the right adrenal metastasis (adrenal2) which was seeded by the left adrenal metastasis (adrenal1; Figure 5-23) was detected 139 days after surgery and after several cycles of chemotherapy were administered (Figure 5-22). Analyzing the mutational processes active in the last shared clone between the two adrenal metastases showed no evidence of a platinum chemotherapy signature meaning that divergence of these two metastases must have occurred prior to the patient receiving treatment. This again highlights early presence of the left adrenal metastasis. Furthermore, the timing of divergence of the two metastases gives an indication to when the cells of the right adrenal metastasis first existed which could be used to infer the growth rate of this metastatic site.

Finally, none of the scans taken before tissue harvest were able to detect the liver metastases that were subsequently sampled. This could either indicate micro-metastatic disease in the liver that could not be detected on imaging, or a very late and rapid expansion of a minor clone from the primary tumor. The timing of the emergence of the liver metastases is not possible given the current information. However, other sources, such as ctDNA could help shed light on the presence of the metastatic clone seeding the liver.

5.4.6 Characterizing uncertainty in migration patterns

In some cases, when running MACHINA under the different hypotheses (single source, multi-source, reseeding), different migration graphs were obtained. As MACHINA infers the most parsimonious migration patterns, this means that multiple solutions are equally likely. Given the current implementation of MACHINA, obtaining all possible optimal solutions is not possible, however in a first step the results from the different hypotheses could be used as an approximation. Additionally, for some PEACE cases, the tree building algorithm inferred multiple possible tree solutions which could be used as input to MACHINA for further exploration of the possible migration patterns.

For PEA043, the tree building algorithm inferred a total of 6 possible tree phylogenies and MACHINA was run using all possible hypotheses for all 6 trees leading to a total of 24 solutions. For all 6 trees, MACHINA inferred single source seeding with 14 migrations,
12 co-migrations from a total of 3 or 4 sources. However, the migration graphs differed across the various solutions. Therefore, to explore the multiple solutions the seeding from the primary tumor alone was excluded and 18 possibilities remained.

In total, MACHINA detected 4 unique solutions. The most frequent solution was found 10/18 times (55.56%) and is shown in Figure 5-24. The second most frequent solution was found 6/18 times (33.33%; Figure 5-25) and the patterns are broadly similar to the first solution, only with lymphnode1 and lymphnode2 swapped. The final two solutions (Figure 5-26; Figure 5-27) were both only found once, respectively and were the only solutions with a total of 4 sources.

![Figure 5-24 First possible migration pattern inferred by MACHINA for PEA043.](image1)

![Figure 5-25 Second possible migration pattern inferred by MACHINA for PEA043.](image2)
Although overall, 4 main migration patterns were found, some individual migrations were more consistent across multiple solutions. For this, individual migration paths were considered. For example, in all solutions the primary tumor seeded one progression sample (progression 1; station 4R lymph node). This sample went on to seed a second progression sample (progression 2) also taken from a station 4R lymph node, as well as the initially sampled recurrence (relapse1) which was also taken from a station 4R lymph node. Furthermore, in addition to lymphnode2 seeding all liver samples in the most frequently found solution (Figure 5-24), in the third solution (Figure 5-26) lymphnode2 was found to seed liver1, 2, 3, and 5, giving additional confidence to those specific
migration paths. However, in order to fully explore the confidence in each solution, the complete solution space must be explored.

5.5 Conclusions

The results in this chapter show that the majority of metastases diverge late, after the last clonal sweep in the primary tumor with no difference observed between the timing of divergence and histologies or types of metastases. Similar results have been found by previous studies (Yachida et al., 2010; Yates et al., 2017; Noorani et al., 2020). In some cases, modelling suggested that metastatic potential was acquired five years after the original cells existed (Yachida et al., 2010). Similarly, (Yates et al., 2017) found that metastatic divergence occurred on average at 87% of the molecular age of the primary tumor. While other studies have found predominantly early divergence (Faltas et al., 2016; Brown et al., 2017; Hu et al., 2019), this is likely confounded by under-sampling of the primary tumor. Indeed, (Hu et al., 2019) estimated that dissemination occurred early for 18 out of 23 metastases (from 17 patients), however, three out of four tumors estimated to disseminate late (4 out of 5 metastases), had multi-region sequencing of the primary tumor compared to 6 out of 17 tumors estimated to diverge early. This is further highlighted in the results of this chapter. When considering only late divergence cases from the cohort and down sampling the primary tumor regions to estimate the timing of divergence, a significant number of cases were mis-classified as early divergence. Specifically, when down sampling to a single primary tumor region and estimating the timing of divergence 78% of cases were mis-classified as early.

Although relative timing could be inferred across the cohort, this does not give an indication of when the metastases diverged and disseminated in absolute time. Previous studies have implemented modeling to further infer the timing of metastatic divergence and have found that a significant proportion of early divergence cases likely occur when the tumor was below the limit of CT detection (Tan et al., 2016; Zhao et al., 2016; Hu et al., 2019). These results highlight limitations of screening approaches to detect primary tumors prior to metastatic divergence and dissemination. Nonetheless, in a few example cases, more specific timing of divergence was possible due to imaging and treatment timelines.
Furthermore, the extensively sampled PEACE cohort was introduced, and a genomic overview shown. Using the detailed clinical annotation and imaging it was possible to infer timing of individual migrations within a single patient. Once the full imaging data will be included, it will be possible to extend this timing of metastatic divergence to the full cohort.

Finally, using a published tool MACHINA (El-Kebir, Satas and Raphael, 2018), migration patterns, including metastasis-to-metastasis seeding were characterized within TRACERx and PEACE. Within the TRACERx cohort the role of primary LN metastasis in seeding recurrences was investigated, and primary LN metastases were shown to seed subsequent recurrences in less than 20% of cases. Consistent results have been shown previously in breast, esophageal, prostate, colorectal, as well as lung cancer (Haffner et al., 2013; Naxerova et al., 2017; Ullah et al., 2018; Noorani et al., 2020; Tang et al., 2021). This highlights that primary LN metastases likely represent metastatic potential of the primary tumor rather than a gateway to further metastases. Furthermore, migration patterns were explored more extensively in the PEACE cohort. Although the majority of sites were found to have been seeded directly from the primary, a large subset of metastases were seeded by lymph node, lung, and liver metastases sampled at tissue harvest. While lymph node and lung metastases seeded metastases across multiple tissue types, liver metastases generally seeded other liver metastases. This could indicate that metastases in certain organs in the body are more likely to contribute to wide-spread disease, while others might only lead to increased mutation burden within the same organ. However, to better explore this, a larger cohort will be needed to further quantify the migration patterns.
Chapter 6 Evolution of metastatic disease in pancreatic ductal adenocarcinoma

6.1 Introduction

The previous chapters of this thesis have mainly explored evolution within the primary tumor and differences between the primary tumor and subsequent metastases. In this chapter, I will focus on tumor evolution in late-stage disease, investigating how cells continue to change after metastatic disease has been diagnosed as well as changes arising due to treatment.

For this, a cohort of pancreatic ductal adenocarcinomas is introduced where metastatic biopsies were performed, and the tumors subsequently tracked using cell-free DNA (cfDNA) obtained from plasma. Pancreatic cancer is a very aggressive form of cancer, which is the 5th most common cause of cancer-related death in the United Kingdom, although it is only the 10th most common cancer type (Pancreatic Cancer Statistics, no date). Patients diagnosed with pancreatic cancer have very poor prognosis, with only 5% of patients surviving more than 10 years which is mostly due to pancreatic cancers frequently being diagnosed at later-stages of disease. This highlights the need for a better understanding of the metastatic process in pancreatic cancer and identification of new potential targets for therapy.

Previously pancreatic cancer therapies have been limited, however more recent studies have explored survival benefits associated with targeted therapies, such as PARP inhibitors (Golan et al., 2014, 2019; Pishvaian et al., 2019). Within the cohort presented in this chapter, a significant subset (21%) of patients presented with germline mutations in either BRCA1 or BRCA2. Therefore, genomic phenotypes could be compared between patients harboring a germline BRCA variant and patients with wildtype BRCA, and subsequently similar patterns were investigated within the Cancer Genome Atlas (TCGA) pancreatic ductal adenocarcinoma cohort (Cancer Genome Atlas Research Network, 2017).

Two models of pancreatic cancer evolution have been proposed: initially, a step-wise model was proposed which highlighted a continuous accumulation of alterations that
eventually lead to an invasive cancer (Hruban et al., 2000; Reiter and Iacobuzio-Donahue, 2016); more recently, a punctuated model was described which is based on the observation that driver alterations can occur simultaneously through complex rearrangements (Notta et al., 2016). Furthermore, a high mutational conservation has been described between the primary tumor and metastasis (Campbell et al., 2010; Yachida et al., 2010; Makohon-Moore et al., 2017; Connor et al., 2019). Although greater genomic instability and polyploidy have been observed in metastases (Connor et al., 2019; Chan-Seng-Yue et al., 2020), the copy number landscape after metastasis has not been extensively explored. In this chapter, the data from cfDNA samples collected after the metastasis biopsy can help give insight into ongoing copy number heterogeneity.

The work described in the following chapter is part of a manuscript currently under review. The data were collected by collaborators as part of a clinical study. The adaptation of the pipeline and processing of the data, as well as all analyses were performed by me.

6.2 Cohort overview

6.2.1 Timelines

The timelines of a cohort of pancreatic cancer metastases and cfDNA samples, including date of diagnosis, metastasis biopsy, as well as when the cfDNA samples were collected for each patient are highlighted in Figure 6-1. The analyzed cohort consisted of 24 patients of which one patient was lost to follow up (PAN117) while three were still alive when the cohort was locked down at the end of 2018 (PAN122, PANVH2, PANVH4). For the remaining patients, the mean overall survival was 424.35 days with a range between 25 days and 1734 days. The date of diagnosis ranged from November 2012 to December 2017.

Out of the 24 patients, only two patients did not have a metastasis biopsy sequenced (PAN106, PANVH2), while for one patient in addition to the metastasis two regions of the primary tumor were sequenced (PAN103). For eight patients (PAN103, PAN104, PAN105, PAN108, PAN109, PAN110, PANVH1, PANVH2), patient derived xenograft (PDX) models were derived from the metastasis biopsy and one PDX sample per patient (where such a model was derived) was sequenced as well. The number of cfDNA
samples ranged from 1 to 8 per patient with a mean of 2.29 samples per patient. Overall, 55 cfDNA samples were sequenced and in total 87 tumor samples were subject to whole exome sequencing for the 24 patients (mean: 3.63, range 1 – 12).

![Image](image-url)

**Figure 6-1 Timelines of patients from diagnosis and sample collection to death.**

The orange dots highlight how many days post diagnosis a metastatic biopsy was obtained. The smaller blue and gray dots show the time points at which cfDNA was sampled. Gray dots indicate that no tumor was detected in the sample, while the different levels of blue show the tumor content in the cfDNA. Finally, the green lines indicate when the patients received systemic anti-cancer therapy.

### 6.2.2 Genomic overview

All obtained samples were aligned to the reference human genome (hg19) using Burrows-Wheeler Algorithm (BWA) (Li and Durbin, 2009).

The median coverage of the tumor samples (excluding the PDX samples) was 103 (range 38 – 201), while the median coverage of the germline blood samples was slightly lower (median of 94).
As the PDX samples could include mouse contamination, the alignment step of the bioinformatics pipeline was adapted to account for this and to filter out any potential mouse reads. In short, the approach applied, classifies each individual read as human or mouse and only the human reads were included in the subsequent pipeline run as well as any analyses. More specifically, first the PDX samples were aligned to the reference human genome (hg19) as well as to the reference mouse genome (mm10) using BWA. Subsequently, bamcmp (Khandelwal et al., 2017) was used to categorize reads into human only, human better, mouse only, and mouse better. Bamcmp classifies each read independently and in the first instance determines whether a read aligns only to human or mouse and therefore classifies it as human only or mouse only respectively. If a read can be aligned to both reference genomes, bamcmp uses the alignment scores to determine whether the read maps better to human or mouse and classifies it as human better or mouse better accordingly. As the raw data is aligned to the human and mouse genomes using the same alignment tool, the scores are directly comparable. Finally, all reads mapping better to the human reference (human only and human better) were combined and used in downstream processing and analysis.

The samples were processed as described in (Jamal-Hanjani et al., 2017), but some filters to infer true positive somatic mutations were adapted, as described below. VarScan2 (Koboldt et al., 2012) and MuTect (Cibulskis et al., 2013) were used to identify somatic variants. A mutation was passed if the variant allele frequency (VAF) was greater than 2% and was called by both MuTect and VarScan2. If the mutation was only called by VarScan2 a VAF of 5% was required. In both cases, a somatic p-value of <= 0.01 was required for the mutation to be called by VarScan2. Furthermore, a sequencing depth >= 10 was required in every region. More than 5 reads were needed to support the variant call, however, if there were >= 3 reads supporting the variant call and the mutation was called by both VarScan2 and MuTect, the mutation was passed even with these lower filters. Finally, the number of reads supporting the variant in the germline needed to be less than 3 with a VAF <= 1%.

Purity and ploidy solutions were obtained from ASCAT (Van Loo et al., 2010) and manually inspected. Of the cfDNA samples, 6 samples had sufficient tumor purity to estimate copy number aberrations while the remaining 49 had very low tumor purity (less than 5%). Furthermore, 6 metastasis biopsies as well as one region of the primary tumor obtained from PAN103 had less than 5% tumor content. The remaining 16 metastasis
biopsies, the remaining region of the primary tumor from PAN103, as well as the 8 PDX samples had sufficiently high tumor content (greater than 15%). The mean purity of these 31 samples was 54% (range 15 – 100%).

For two patients (PAN114, PAN122) no mutations passed the filtering described above, therefore these patients were excluded from subsequent analyses. Additionally, three patients (PAN106, PAN119, PAN121) had no samples with sufficient tumor purity to determine copy number status, however mutation data was still available for these cases. Finally, for eight patients (PAN103, PAN104, PAN105, PAN108, PAN109, PAN113, PAN125, PANVH1) multi-region copy number data was available.
Figure 6-2 Genomic overview of cohort.

The top panel shows the mutation load for each patient split into clonal mutations, that are found in all samples, and subclonal, mutations found only in some regions. The second panel shows the proportion of mutational signatures for each patient. Five patients (on the right, grayed out), did not have enough mutations (less than 50) to perform reliable signature deconvolution. The bottom panel shows different genomic events for a set of cancer genes of interest. Mutations are represented as squares and copy number events are overlayed as triangles. The color of the tiles and triangles represents the timing (clonal meaning
that a mutation was found in all tumor cells and subclonal meaning that the mutation was detected in a subset of cancer cells) of the mutations and copy number events respectively. Patients for which no copy number information is available are highlighted in gray at the bottom.

The median number of mutations within the cohort of 22 patients included for further analysis was 84 mutations (range 9 – 718) with a median clonal mutation load of 69 mutations (range 9 – 178) and a median subclonal mutation load of 12 mutations (range 0 – 587) (Figure 6-2). The proportion of subclonal mutations ranged between 0% and 82%.

To understand the mutational processes at play during tumor evolution mutational signature analysis was performed. Mutational signatures represent various exogenous and endogenous mutagenic processes which lead to single base substitution (SBS) as well as doublet base substitutions (DBS), small insertions and deletions (ID), copy number aberrations and structural variants (Nik-Zainal et al., 2016; Alexandrov et al., 2020; Drews et al., 2022; Steele et al., 2022). For the present analysis only SBS signatures were considered, and the presence of a set of selected signatures which are frequently found in pancreatic cancer was explored (Alexandrov et al., 2013; Aguirre et al., 2018). The set of signatures includes SBS1 which is initiated by spontaneous deamination of 5-methylcytosine, as well as SBS5 (etiology unknown), which are both considered clock-like signatures which correlate with the age of the patient in most cancer types. Furthermore, SBS2 and SBS13 were included which are both linked to activity of the APOBEC family of cytidine deaminases, and have previously been found in pancreatic adenocarcinoma (Nik-Zainal, Alexandrov, et al., 2012; Alexandrov et al., 2013). Finally, SBS3 and SBS6 were included which are linked to defective homologous recombination-based DNA damage repair and defective DNA mismatch repair respectively. Both of these signatures have also been described previously in pancreatic adenocarcinoma (Alexandrov et al., 2013).

The signature analysis was performed using the R package deconstructSigs (Rosenthal et al., 2016) and the deconvolution was implemented both for all mutations (Figure 6-2) as well as by separating clonal and subclonal mutations (Figure 6-3). In most cases (18 of 22) there were insufficient subclonal mutations to perform signature deconvolution, and additionally in seven cases there were also insufficient clonal mutations.
Previous studies have suggested that BRCA-like tumors, including tumors with active SBS3 signature, could respond well to PARP inhibitors (Stok et al., 2021). Therefore, the extent of tumors with active SBS3 was explored within the cohort. In 13 of 22 cases, SBS3, which is linked to defective homologous recombination-based DNA damage repair, was detected as an active signature when considering all mutations (overall at least 50 mutations; weight of signature >= 0.06 and >= 10 mutations associated with the signature). When considering clonal and subclonal mutations separately, SBS3 was active in 12 patients. For one of these patients SBS3 had not passed the filters when considering all mutations (PAN105). Additionally, for two patients (PAN109, PAN113) where SBS3 was active overall, no SBS3 activity could be detected in the clonal mutations. For one of these patients (PAN113), SBS3 was active in the subclonal mutations only and additionally, in one patient (PANVH1) SBS3 was active both clonally and subclonally.

![Figure 6-3 Overview of mutational signature proportions for the cohort separated by clonality. Patients are ordered by the proportion of SBS3 found in the clonal mutations. If insufficient mutations were detected (less than 50 mutations) to perform reliable signature deconvolution, the bars are shown in faded colors. The fraction of mutations which could not be attributed to any of the analyzed signatures (other) are shown in gray.](image)

Finally, genomic events in a set of cancer genes were analyzed. The gene list consists of five genes frequently mutated in pancreatic cancer (KRAS, TP53, CDKN2A, SMAD4,
Chapter 6 Results

*TGFBR2* (Cancer Genome Atlas Research Network, 2017; Bailey et al., 2018), as well as genes involved in homologous recombination repair and DNA damage repair (*PALB2, ATM, BRCA1, BRCA2*). Variants in genes linked to the DNA damage repair pathway have been shown to contribute to a BRCA-like phenotype of tumors (Stok et al., 2021). A set of 11 genes were also included where at least one mutation found in the cohort was classified as a driver mutation (see Section 2.2.1.8). Mutations in this list of a total of 21 genes were classified as occurring clonally or subclonally and are shown in Figure 6-2 as colored tiles.

For seven patients initially no *KRAS* mutations passed the filtering described above. In three cases this was due to low coverage in one of the regions sequenced, which was subsequently excluded from analysis, and therefore the mutations then passed the filtering. In four cases this was due to higher VarScan2 somatic p-values than the filter allowed. However, due to the low tumor purity in some of these samples and the fact that all *KRAS* mutations occurred in known hotspots, these were subsequently rescued and included in the analysis. Three of these filtered mutations were G12D and three G12V mutations while one was a Q61H mutation.

For the same set of 21 genes, copy number events were also analyzed. For this, copy number data for each segment in each sample was divided by the calculated mean ploidy of the sample and log2 transformed. Gains were defined as ratios greater or equal to \(\log_2(2.5/2)\) and losses were defined as ratios less or equal to \(\log_2(1.5/2)\), or loss of heterozygosity (LOH) events where the minor allele had a copy number state of 0. Such gains and losses for the genes of interest are shown in Figure 6-2 as triangles. For patients where only a single sample passed copy number estimation, all events were classified as clonal. However, for patients with multiple samples where ploidy could be estimated, copy number events were classified as clonal if they were present in all samples. For these patients, events were classified as subclonal if only found in a subset of regions, or if there were opposite classifications in different regions. In this case, when summarizing all events for Figure 6-2, only losses were considered in tumor suppressor genes while gains were considered in oncogenes, as these events could represent putative copy number driver alterations. The individual copy number events for each sample of a patient are highlighted in Figure 6-4. In one patient, PAN105 both a clonal gain and loss of *HNF1A* were observed (Figure 6-4). This was due to the gene spanning two segments, the first of which was classified as an LOH (copy number states of 1 and 0 of the two alleles), while the second was classified as a gain event (copy number states
of 2 and 1 of the two alleles). This was observed in both the metastasis biopsy as well as the PDX sample of this tumor.

For the three patients where no samples passed copy number estimation (highlighted in gray at the bottom of Figure 6-2), only mutation events are shown.

The results show similar genomic patterns as have previously been described in pancreatic ductal adenocarcinoma (Cancer Genome Atlas Research Network, 2017). Most of the tumors within the cohort had a KRAS mutation, as well as a clonal mutation or clonal loss of TP53. Additionally, mutations in other cancer genes were observed, however only in a small number of patients. This could indicate that early pancreatic evolution follows a relatively deterministic evolutionary pattern with low amount of variability related to mutations. Similarly, when investigating copy number events occurring in cancer genes of interest, not much heterogeneity is observed, with loss of
TP53, CDKN2A and SMAD4 occurring clonally in a large subset of patients. Although the majority of mutation and copy number drivers are consistent across the cohort, the mutational burden of the patients shows higher variation. This indicates that other factors might be contributing to genomic heterogeneity between tumors.

6.3 Comparing BRCA versus wildtype patients

To date therapies for pancreatic cancer are limited. For patients presenting with metastatic disease, the chemotherapy agents FOLFIRINOX and gemcitabine with nab-paclitaxel are the main treatment strategies (Sohal et al., 2020). However, even with treatment, median overall survival times are less than 12 months (6.7-6.8 months with gemcitabine alone, 8.5 months with gemcitabine and nab-paclitaxel, 11.1 months with FOLFIRINOX) (Conroy et al., 2011; Von Hoff et al., 2013).

More recently, novel targeted therapies have been explored. Multiple studies have shown increased overall survival in patients with germline BRCA1 or BRCA2 mutations treated with PARP inhibitors due to synthetic lethality by inactivating both PARP1 and BRCA (Golan et al., 2014, 2019; Pishvaian et al., 2019). Whether these treatments will also benefit patients presenting only with somatic mutations in BRCA1/2 remains to be determined (Lowery et al., 2017; Brown and Reiss, 2021).

Within the cohort, five (21%) patients were identified with germline alterations in BRCA1 or BRCA2. Additionally, six (25%) patients had somatic mutations or loss of heterozygosity in at least one of the BRCA genes, in five out of six cases these events were clonal. Overall, only three patients had both a germline and somatic alteration leading to loss of function.

To explore whether germline and somatic BRCA alterations may result in a distinct genomic phenotype, in this section differences in genomic and clinical characteristics were explored between patients with germline (GL) and somatic alterations in BRCA1/2 as well as patients with BRCA wildtype (WT).

6.3.1 Mutation level comparison of BRCA versus wildtype

Tumors in patients with a germline BRCA1/2 alteration presented with a trend for higher mutation burden compared to tumors in patients with wild type BRCA1/2 (median
mutation burden 201 vs 68 respectively; Wilcoxon rank sum test, p = 0.16; Figure 6-5). When splitting the mutations into clonal and subclonal mutations, the trend remained (Clonal: median mutation burden 115 vs 68 in BRCA GL vs WT cases; Wilcoxon rank sum test, p = 0.089; Subclonal: median mutation burden 170 vs 16 in BRCA GL vs WT cases; Wilcoxon rank sum test, p = 0.39; Figure 6-5). When comparing tumors of patients with somatic BRCA alterations to those with GL alterations and tumors with no alterations in BRCA1/2, no significant difference was observed (Wilcoxon rank sum test, p = 0.33 and p = 0.22 respectively; Figure 6-5). This suggests that, although not likely a tumor initiating event, a BRCA germline alteration may potentially lead to more genomically unstable tumors with a higher mutation burden. However, due to the size of the cohort, the results are underpowered.

Both BRCA1 and BRCA2 are involved in multiple cellular processes, most importantly maintaining genomic integrity through homologous recombination (HR)-mediated double strand break repair (Lord and Ashworth, 2012; Davies et al., 2017). Therefore, mutational signature analysis was performed specifically including single base substitution (SBS) signature 3 which has been linked to defective HR DNA damage repair deficiency (Alexandrov et al., 2020) (see Section 6.2.2).
Chapter 6 Results

Tumors in patients with germline mutations in BRCA1/2 had a higher proportion of SBS3 compared to tumors with WT BRCA1/2 (median proportion of SBS3 0.486 vs 0.311 respectively; Wilcoxon rank sum test, p = 0.053; Figure 6-6). This trend was also notable when comparing number of mutations associated with SBS3 in tumors with BRCA GL and WT (median number of mutations attributed to SBS3 137 vs 22.5 respectively; Wilcoxon rank sum test, p = 0.069; Figure 6-6). No significant difference in proportion or number of mutations associated with SBS3 was observed when comparing tumors with somatic BRCA1/2 events to tumors with WT BRCA (Figure 6-6). This could indicate that a somatic BRCA alteration, at least a single LOH event, is not sufficient to drive a BRCA-like phenotype.

![Figure 6-6 Comparison of proportion and mutation counts of SBS3 in patients with and without BRCA alterations.](image)

Patients were classified into three groups: patients with a GL BRCA alteration, three of which also additionally had a somatic mutation or loss or LOH event; patients with a somatic BRCA mutation, or loss or LOH event; patients with no BRCA alterations. The left panel shows the proportion of SBS3 detected in the signature deconvolution while the right panel shows the number of mutations associated with SBS3.

### 6.3.2 Copy number level comparison of BRCA versus wildtype

To compare differences at the copy number level between patients with germline BRCA1/2 alterations to those with somatic BRCA1/2 alterations and those with WT BRCA, the weighted genomic instability index (wGII) (Burrell et al., 2013) was used. The
wGII score was calculated as the proportion of the genome with aberrant copy number relative to the median ploidy, weighted on a per chromosome basis.

The wGII score was slightly higher in tumors with somatic BRCA alterations (median wGII 0.362 vs 0.489 in GL BRCA vs somatic BRCA; Wilcoxon rank sum test, p = 0.61; Figure 6-7). Similarly, the wGII score was also slightly higher in tumors with somatic BRCA1/2 alterations compared to WT BRCA tumors (median wGII 0.489 vs 0.238 in somatic BRCA vs WT BRCA; Wilcoxon rank sum test, p = 0.27; Figure 6-7).

As another measure to compare the copy number, the proportion of LOH identified across the genome was calculated for all patients. The proportion of LOH was significantly higher in tumors with somatic BRCA alterations compared to WT BRCA tumors (median proportion LOH 0.418 vs 0.176 in somatic BRCA vs WT BRCA; Wilcoxon rank sum test, p = 0.0496) and a trend towards significance between tumors with somatic BRCA alterations compared to tumors with GL BRCA alterations was observed (median proportion LOH 0.418 vs 0.266 in somatic BRCA vs GL BRCA; Wilcoxon rank sum test, p = 0.17; Figure 6-8).
The majority of somatic BRCA1/2 alterations classified in the cohort are loss of heterozygosity at the BRCA1 or BRCA2 locus. Therefore, it is unclear whether the significantly higher proportion of LOH is driven by LOH of the BRCA1/2 locus, or whether the higher proportion of LOH found in these tumors makes LOH of BRCA1/2 more likely.

Figure 6-8 Comparison of proportion of LOH. Patients were classified into three groups: patients with a GL BRCA alteration, three of which also additionally had a somatic mutation or loss or LOH event; patients with a somatic BRCA mutation, or loss or LOH event; patients with no BRCA alterations. The fraction of LOH (fLOH) is shown for all patients across the three groups.

6.4 Validation of BRCA vs WT in TCGA

Due to the size of the cohort presented here and the trend seen in comparisons of GL BRCA vs WT BRCA patients in relation to the mutation burden as well as the prevalence of SBS3, a validation using TCGA data was performed (Cancer Genome Atlas Research Network, 2017). For this, all pancreatic tumors from the TCGA cohort were considered that also had germline and clinical information available. The final cohort used in the validation consisted of 140 patients with pancreatic cancer.

Possible germline BRCA alterations were extracted using ANNOVAR (Wang, Li and Hakonarson, 2010) and reviewed manually to determine patients with known BRCA germline alterations. In the final cohort, five patients (3.6%) with germline alterations in
BRCA1 or BRCA2 were identified. This represents a significantly lower proportion of germline BRCA patients compared to the cohort introduced here (Fisher’s exact test, \( p = 0.01 \)).

### 6.4.1 TCGA overview

![TCGA cohort overview](image)

Figure 6-9 Overview of pancreatic ductal adenocarcinoma TCGA cohort used to validate findings. The patients (shown along the x-axis) are split into those classified as having a GL BRCA alteration (on the left; 5 patients) and those without a GL BRCA variant. The top panel shows the number of mutations detected, the second and third panel show the presence of somatic SNV and copy number LOH, respectively, for a set of 10 genes of interest. The bottom 8 panels show clinical information for each patient, including gender, tumor size, and stage.

### 6.4.2 Mutation level comparison of TCGA BRCA GL vs WT

The mutation burden in tumors with germline BRCA alterations was significantly higher compared to tumors with WT BRCA as well as compared to tumors with somatic BRCA
alterations, which were predominantly LOH of BRCA1/2 in the TCGA cohort (median mutation burden 57 vs 31 in GL BRCA vs WT BRCA; Wilcoxon rank sum test, p = 0.0033; median mutation burden 57 vs 32 in GL BRCA vs somatic BRCA; Wilcoxon rank sum test, p = 0.0055; Figure 6-10). No significant difference in mutation burden was observed between tumors with somatic BRCA and WT BRCA (Wilcoxon rank sum test, p = 0.66).

When comparing the number of mutations associated with SBS3 detected in the TCGA cohort between tumors with GL BRCA and WT BRCA, the number of SBS3 mutations was significantly higher in GL BRCA tumors (median number of SBS3 mutations 20.9 vs 3.89; Wilcoxon rank sum test, p = 0.041; Figure 6-11). Similarly, the number of mutations associated with SBS3 was also higher in GL BRCA compared to somatic BRCA tumors, although not significantly (median number of SBS3 mutations 20.9 vs 3.48; Wilcoxon rank sum test, p = 0.11; Figure 6-11). Finally, no significant difference was observed when comparing the number of SBS3 mutations between tumors with somatic BRCA and WT BRCA (Wilcoxon rank sum test, p = 0.93).

These results are consistent with the trend seen in the pancreatic cohort described above. They suggest that tumors in patients with GL BRCA alterations behave differently.
from tumors with somatic or WT BRCA. Even in cases where only a single germline alteration in BRCA was found, without a second hit somatically, similar patterns could be observed. This suggests that BRCA alterations might not be a founding event of pancreatic ductal adenocarcinoma, however the evolution of tumors with a germline alteration in BRCA follows different patterns than those with somatic BRCA alterations. What the effect of these differences might be on response to treatment remains to be determined, but could indicate that response to targeted treatments, such as PARP inhibitors might be limited when only a somatic BRCA alteration is detected in a tumor.

**Figure 6-11 Comparison of SBS3 in TCGA.**

Similar to above, the patients were classified into three groups: patients with a GL BRCA alteration, three of which also additionally had a somatic mutation or loss or LOH event; patients with a somatic BRCA mutation, or loss or LOH event; patients with no BRCA alterations. The number of mutations associated with SBS3 is shown for the TCGA cohort.

### 6.5 Exploring cfDNA samples

To explore the dynamic evolution of these tumors, cfDNA samples were collected for each patient. From the 24 patients in the cohort, 55 cfDNA samples were sequenced, with at least one cfDNA sample from each patient (median 2, range 1 – 8). As mentioned in Section 6.2.2, two patients were excluded from subsequent analysis due to no mutations being detected. Therefore, the four cfDNA samples from these two patients were also not taken forward for further analysis.
6.5.1 High purity cfDNA samples

Overall, 6 cfDNA samples from 4 patients were estimated to have high tumor content (median purity 0.36, range 0.26 – 0.76). This means that in these cases, 26% to 76% of the DNA found in blood could be attributed to the patients’ tumors. This is significantly higher than previous studies have found where the ctDNA levels ranged from below the limit of detection to approximately 6%, although higher tumor content of up to 90% has been found in some cases (Diaz and Bardelli, 2014; Lipson et al., 2014; Peneder et al., 2021).

For one patient, PAN113, all three cfDNA samples that were sequenced had high purity (0.7, 0.37, and 0.76). For the other three patients with a single high purity cfDNA sample, one had an additional low purity cfDNA sample while the other two tumors had only the single high purity sample. Therefore, it is difficult to determine whether the high purity samples appear as a result of a tumor intrinsic property with all cfDNA samples presenting with high tumor purity or whether this is due to other factors such as tumor lysis syndrome which might occur intermittently (Davidson et al., 2004).

![Figure 6-12 Example copy number profiles. The top panel shows an example of a high purity PDX sample, the middle panel a low purity cfDNA sample and the bottom panel one example of a high purity cfDNA sample.](image)
6.5.2 Setting filters for calling mutations in cfDNA samples

As mentioned in Section 6.5.1, for a small subset of cfDNA samples (6 of 51), high tumor purity (>= 15%) was observed. However, in the majority of cfDNA samples (45 of 51) tumor purity was below 5%. Therefore, tumor content and copy number aberrations could not be estimated using ASCAT. Due to low tumor content, in most cases de novo mutation calling using the filters described in Section 6.2.2 only yielded a low number of mutations (median number of mutations: 3, range 0 – 100). In seven cfDNA samples no de novo mutations were found while in the majority of samples (41 of 51) less than 10 mutations were called.

In order to increase the number of mutations detected in the cfDNA samples, mutations confidently called in other samples of the same patient were also considered. To reduce the rate of false positive mutations, multiple thresholds were explored.

Initially, all mutations that were confidently called in at least one sample of the patient, based on the filters described in Section 6.2.2, were collated. Mutations in cancer genes that were classified as driver mutations were removed from this analysis as some overlap of such mutations can be expected. Subsequently, for each patient, read information for these high confidence mutations was extracted from the original alignment files of the samples of all other patients using bam-readcount (https://github.com/genome/bam-readcount).

For each patient, mutations in each cfDNA sample were called from the initially passed mutations using a variant read count filter of 1, 2, 3, 5, 10 or more reads and a VAF filter of greater than 0%, 1%, 1.5%. Mutations positively identified in the cfDNA sample of interest based on these filters were subsequently explored in all other cfDNA samples. The proportion of mutations positively identified based on the same filters in at least one other cfDNA sample was calculated and weighed against the total number of mutations identified in the cfDNA sample of interest. The results can be seen in Figure 6-13.

In order to determine a reasonable threshold with which to positively identify high confidence mutations in the cfDNA, the tradeoff between total number of mutations identified and proportion of mutations also found in at least one other cfDNA sample was considered. When using a lenient filter of one or more variant read and a VAF greater than 0%, at least one mutation was identified in all cfDNA samples (median 24; range 1
However, in this case more than 25% of mutations for all cfDNA samples were found in samples of other patients (median 73%; range 33 – 100). When using a very stringent filter of 10 or more variant reads and a VAF greater than 1.5%, 42 cfDNA samples had less than 5 mutations (median 2; range 0 – 107) but in this case only a single mutation was found in another cfDNA sample while the remaining samples did not share mutations.

![Figure 6-13 Overview of effect of different filters on mutation load and false positive rate.](image)

The figure shows the number of mutations called in each cfDNA sample on the y-axis and highlights the proportion of such mutations found in cfDNA samples of other patients on the x-axis. The columns highlight the read count filter used while the rows show the VAF filter used to define a mutation as present. The gray area in each plot highlights samples that have less than 5 mutations called or where more than 25% of mutations are found in samples of other patients.

When optimizing for number of cfDNA samples retained with five or more mutations as well as where less than 25% of such mutations were found in other cfDNA samples, a
variant read count of 3 or more and a VAF of greater than 1% or 1.5% gave the best results (34 and 36 samples retained, respectively; Figure 6-14).

Overall, the median number of mutations called in the cfDNA samples increased to 9 (range 0 – 117) with a median increase of 4 mutations per sample (range 0 – 50). In 18 cfDNA samples, 10 or more additional mutations found at lower frequency could be identified using these updated filters.

![Figure 6-14 Number of cfDNA samples retained for each set of filters.](image)

### 6.5.3 cfDNA samples with signature of unknown origin

In the tumors of two patients within the cohort, a large proportion of mutations could be attributed to SBS6 which is associated with a defective DNA mismatch repair mechanism (Alexandrov et al., 2013; Meier et al., 2018). Upon further investigation, the signature was predominantly found in the combined mutations of the cfDNA samples rather than the metastasis sample of both patients. Overall, no overlap between mutations could be found between the mutations in the cfDNA and metastasis samples of these patients. To rule out a sample swap being the cause of no overlapping mutations, the VAFs of 24 SNPs previously described by Pengelly et al. (Pengelly et al., 2013; Jamal-Hanjani et al., 2017) were compared across samples (Figure 6-15). This confirmed that all samples were truly derived from the same patients, respectively.
In PAN116 the signatures found in the metastasis sample were SBS1 (12.1%) which is a clock-like signature (Nik-Zainal, Alexandrov, et al., 2012; Alexandrov et al., 2013), SBS3 (76.4%) which is likely caused by defective homologous recombination DNA damage repair (Nik-Zainal, Alexandrov, et al., 2012; Alexandrov et al., 2013; Zámborszky et al., 2017), as well as a small proportion of SBS6 (11.5%). Conversely, in the cfDNA sample the predominant signature found was SBS6 (91.1%) and a small proportion of SBS3 (8.9%). Although more mutations were found in the metastasis sample (50 mutations in the metastasis vs 8 and 22 in the two cfDNA samples, respectively), the mutational profile looks sparse with mutations present in many trinucleotide contexts, while for the cfDNA almost all mutations found are C to T mutations at a GpC site (Figure 6-17).
In PAN120, the signatures found in the metastasis sample were SBS1 (24%), SBS3 (20.2%) and SBS5 (48.1%) for which the etiology is unknown (Alexandrov et al., 2013), however has also been classified as a clock-like signature which correlates with the age of the individual. In the cfDNA sample again the predominant signature detected is SBS6 (86.7%) with a small proportion of SBS3 (13.3%) also found. Similar to patient PAN116, in PAN120 the majority of cfDNA specific mutations can also be attributed to C to T mutations at GpC sites (Figure 6-17).

The mutational profiles of the cfDNA samples of both patients were combined and compared to all SBS signatures featured on COSMIC (Alexandrov et al., 2020) using cosine similarities. This was done to rule out bias due to the a-priori selection of signatures used for running deconstructSigs (see Section 6.2.2; see Section 2.2.3). The highest cosine similarity with the mutational profile was observed with SBS6 (cosine similarity = 0.895; Figure 6-16). The other signature with a similarly high cosine similarity was SBS15 (cosine similarity = 0.822), which is also linked with defective DNA mismatch repair (Alexandrov et al., 2020).

![Cosine Similarities of the mutational profile and all SBS signatures featured on COSMIC.](image)

In the remaining cfDNA samples there was no clear evidence of a strong signal from SBS6, however this can be attributed in part to low tumor burden in the samples as in these cases, no new mutations were called in the cfDNA. It is unclear whether these distinctive mutations and signature profiles in the cfDNA samples of these to patients can be attributed to a secondary tumor or whether these mutations might be treatment induced, or an unknown artefact.
To verify whether the signature could be artefactual, the transcription strand bias was considered which has previously been linked with artefact signatures (Costello et al., 2013; Alexandrov et al., 2020). Due to low number of mutations found in the cfDNA samples of both patients, transcription strand bias cannot be ruled out (Figure 6-18).
6.6 Evolution of pancreatic tumors

In order to track the evolution of tumors over time using the cfDNA samples, and to begin linking presence and absence of certain mutations to temporary response or resistance to therapy, circulating tumor DNA (ctDNA) positivity within the cfDNA samples was determined. A cfDNA sample was considered tumor positive if it contained five or more mutations with less than 25% of those also found in other cfDNA samples within other patients, and if additionally, at least one clonal mutation was found in the sample. An overview of tumor positive cfDNA samples and their mean VAF of clonal mutations can be seen in Figure 6-19.
Chapter 6 Results

Figure 6-19 Overview of samples and tumor content over time. All samples are shown on the x-axis in temporal order of sampling. CfDNA samples in which no tumor could be detected are shown in gray while samples where sufficient mutations were detected are shown in colors – green for primary tumor samples, dark blue for metastatic samples and light blue for cfDNA samples. The y-axis shows the mean VAF of the clonal mutations that could be detected in each sample.

For the majority of cfDNA samples (45/51), the tumor burden was too low to accurately estimate purity and ploidy of the samples, which meant that calculation of the cancer cell fraction (CCF) was not possible. In some cases, this likely represented no disease burden being present at the point of sampling, which could further be corroborated by little to no mutations being detected (15/51 samples with less than 5 mutations and purity below 5%). However, for the remaining samples, an adapted estimate of the CCF was used to further track mutations over time.

6.6.1 Adapted CCF estimate and validation

The cancer cell fraction (CCF) of a mutation is calculated by adjusting the variant allele frequency by the copy number at the location of the mutation and correcting for the purity of the sample (Jamal-Hanjani et al., 2017) (see Section 2.2.1.6). As the majority of cfDNA samples had low tumor content, an adapted approach was used to calculate an approximation of the CCF. Three cfDNA samples with high purity from one patient were used to validate the approach by comparing the adapted estimates with the CCF of the mutations.
As seen in the top panel of Figure 6-20, just using the VAF multiplied by 2 and capping values at 1 as an approximation for the CCF leads to underestimation of the true CCF values. This approach to approximating the CCF of each mutation essentially assumes that the tumor sample is 100% pure and diploid. The former of these assumptions is clearly not correct, with the majority of samples having a significantly lower purity, therefore leading to significant underestimation of the true CCF values.

To improve the adjusted CCF value, an approximation of the purity of the sample was incorporated by dividing the VAF of each mutation by the median VAF of all clonal mutations. This value was subsequently capped at 1 (Figure 6-20 mid panel). By incorporating this correction factor for purity, the estimates of the adapted CCF are greatly improved compared to using the VAF alone with both the correlation coefficient (R) and the slope of the fitted linear model increasing.

In the final adjustment of the method, the copy number of any high purity sample was used to correct the estimate further. In this case, the greatest mutation copy number (see Section 2.2.1.6) across all high purity samples was selected for each mutation. If this mutation copy number was below 1, i.e., the mutation was either acquired subclonally or lost in a subset of cells, no correction was performed. For all mutation copy numbers greater than 1, the VAF of each mutation was first divided by the mutation copy number and then divided by the median such value across all clonal mutations, as described above. The values were again capped at 1 (Figure 6-20 bottom panel). This additional adjustment yielded slightly lower estimates of the slope of the linear model and the correlation coefficient decreased in one sample. However, the sum of distances of each point to the identity line decreased strictly from the adjusted VAF to the final adjustment of the CCF across all three samples.
Figure 6-20 Validation of adapted CCF calculation for three samples of one patient. The three samples are shown in the three columns of the plot. The x-axes show the true CCF value calculated for each mutation for each of the three samples. In the top row, the adjusted VAF (adjVAF) is shown as an approximation to the CCF by multiplying the observed VAF by 2. The adjusted CCF shown in the middle row (adjCCF1) corrects for tumor purity by using the median VAF of the clonal mutations. The adjusted CCF shown in the bottom row (adjCCF2) additionally also corrects the VAF for any gains found in the high purity sample.

6.6.2 Determining copy number aberrations through phasing

In the majority of cfDNA samples (45/51) the tumor purity was lower than 5% meaning no copy number aberrations could be detected using ASCAT. However, in most patients (19/22) at least one tumor sample (metastasis biopsy or PDX) with purity greater than
5% was sequenced and could be leveraged for estimation of copy number aberrations. In a given patient, the segmentation as provided by ASCAT for each of the samples was overlapped to determine the minimum consistent regions across all samples. Each of these regions had a number of SNPs associated with the segment, with individual BAF estimates for each SNP as well as a segmentBAF calculated by ASCAT.

For each of these minimum consistent regions, the sample with the largest segmentBAF was selected and all SNPs associated with the segment were classified as belonging to allele A or B based on the individual BAF estimates within that sample being greater or less than 0.5, respectively. The classification of SNPs was then applied to all other samples in a given patient. If the segmentBAF of all samples was equal to 0.5, meaning no copy number aberration was detected at that locus, the SNPs were classified as greater or less than 0.5 randomly across all samples. In this manner a false phasing, i.e., illusion of phasing, could be avoided.

To determine whether a copy number aberration was present in another sample of the same patient, multiple comparisons were performed to test whether the distribution of SNPs from allele A and B were significantly different in the sample of interest. For this comparison, only segments were considered where a copy number aberration was detected by ASCAT in at least one region. A Wilcoxon test was used to compare the BAF distributions of the SNPs belonging to allele A and B ($p < 0.001$). Additionally, for each SNP on allele A, a paired BAF could be estimated for allele B by calculating the difference of the original BAF to 0.5. In this manner, the number of SNPs could be doubled, and a paired t-test could be performed to determine whether the distributions of BAFs of allele A and B differed significantly ($p < 0.001$). Finally, a Wilcoxon test was used to compare BAF distributions within the germline sample of the patient to account for somatic copy number aberrations only ($p > 0.1$). If all three conditions were met, a copy number aberration was called in a given segment and sample.

Finally, for segments with copy number aberrations, mirrored subclonal allelic imbalance (MSAI) (Jamal-Hanjani et al., 2017) was determined. If for a given segment, the SNPs
classified as belonging to allele A, i.e., the major allele, in one region were found to have a mean BAF below 0.5 in another region, this segment was defined as MSAI.

### 6.6.3 Examples of evolution over time within pancreatic ductal adenocarcinoma

In this section, the tools described above were applied to investigate the evolution of pancreatic tumors over time using cfDNA. For this, the adapted CCF estimate, as described in the section above (adjCCF2), was used as a proxy for the CCF of mutations in samples where no CCF could be estimated directly, otherwise the true CCF estimate was used. For a subset of patients with at least two high purity samples, phylogenetic trees could be reconstructed. An overview of four patients with samples from multiple time points, as well as phylogenetic trees are shown in Figure 6-21. The plots highlight that in some cases, cfDNA specific mutations were found, while other mutations found in the metastasis sample were not detected in the cfDNA. This could indicate that mutations in certain cell populations do not confer a fitness advantage to survive in the bloodstream, or that these cells were sensitive to treatments that might have been administered during this time and were therefore not detected in subsequent samples. Several examples are highlighted in more detail in the next sections.
Figure 6-21 Evolution over time of pancreatic tumors. For each of the four patients shown, the figure highlights the time at which the samples were obtained. The top panel shows the mean variant allele frequency of the truncal cluster as an approximation of tumor burden. The bottom panel shows the cancer cell fraction estimates for all clusters detected in the tumor at each timepoint. On the right side the evolutionary tree and the cluster relationships are highlighted while at the top, the branches of the tree are highlighted that are detected in the different samples.

6.6.4 Patient PAN103

Patient PAN103 initially presented with a primary pancreatic tumor in January 2013. This tumor was resected surgically and subsequently the patient received first line treatment. In November 2013 the tumor progressed but responded well to the second line treatment given from November 2013 to May 2016. At this stage the tumor progressed again, and the liver metastasis was biopsied and cells from the biopsy were used to derive a PDX model of the metastatic tumor. Subsequently, the patient had 8 blood samples taken for cfDNA sequencing (Figure 6-22).
In the majority of cfDNA samples (6/8) no mutations were detected. However, in two samples from February and June 2017, 2 and 3 clonal mutations, respectively, were detected.

The metastatic sample also had a very low tumor burden (less than 5% tumor content and no clonal mutations detected) with the majority of the sample likely containing stroma. However, using the primary tumor sample and the PDX a phylogenetic tree could be reconstructed (Figure 6-23).

The patient was found to have a germline BRCA2 alteration, a clonal KRAS mutation, as well as an additional subclonal mutation in BRCA2 that was found solely in the primary tumor and not the PDX sample. This highlights that the somatic BRCA2 alteration likely was not responsible for tumor initiation as it was not present in all cancer cells. Additionally, the fact that the mutation was present in the primary tumor and not present in the PDX might suggest that cells not harboring this somatic BRCA2 alteration were selected for due to the treatment given.

*Figure 6-22 Timeline of PAN103 showing treatment overview as well as sample collection.*
6.6.5 Patient PAN113

Patient PAN113 was diagnosed with metastatic pancreatic ductal adenocarcinoma in early 2017. A metastasis biopsy was performed 31 days later at which point a cfDNA sample was also extracted. Additional blood samples for cfDNA were obtained 100 and 157 days post diagnosis (Figure 6-24). Due to copy number estimates being available for the metastasis sample as well as all three cfDNA samples, an adapted version of PyClone (Roth et al., 2014; Jamal-Hanjani et al., 2017) was used to cluster mutations based on their CCF.
The timeline highlights when the patient was diagnosed and how many days post diagnosis the metastatic biopsy was sampled. Furthermore, it shows the times at which cfDNA samples were obtained and overlays the timings when systemic anti-cancer therapy was administered.

In the initial metastatic sample only two distinct clusters of mutations were found (cluster 1 & 4), one of which is the truncal cluster (Figure 6-25; Figure 6-26). In the initial cfDNA sample the subclonal metastatic cluster could not be detected, however three additional clusters could be determined (clusters 2, 3, 5).

![Phylogenetic tree of PAN113](image)

*Figure 6-25 Phylogenetic tree of PAN113.*

*The driver mutations associated with different clusters are annotated and clusters emerging during treatment are also highlighted.*

In between day 50 and day 137 the patient received first line treatment of Gemcitabine and nab-paclitaxel. During this treatment the second cfDNA sample was taken. In this sample a new clone emerged (cluster 6) while one of the clones present in the initial cfDNA sample started decreasing in size (cluster 2). This could be due to sensitivity to the treatment given. However, the remaining clones found in the first cfDNA sample were also found in the second sample highlighting the heterogeneity found in the tumor, and also in relation to treatment response. In the final cfDNA sample the same clones remained present with an additional clone rapidly emerging at this time (cluster 7). This emerging clone might be due to selection pressures changing post-treatment, which was
further supported by a driver mutation in \textit{CREBBP} present in the cluster (Figure 6-25; Figure 6-26).

![Figure 6-26 Cancer cell fraction of clusters in PAN113 over time.](image)

In addition to the clonal dynamics in this patient, parallel events were observed at the copy number level (segments defined as MSAI). Although both the metastasis sample and the cfDNA samples showed copy number events on chromosome 4q and 18p, these events occurred on different alleles (Figure 6-27).
Figure 6-27 Copy number profiles of metastasis and three cfDNA samples of patient PAN113.

For each segment, a region was selected that had the greatest absolute difference of BAF values to 0.5. The SNPs were then colored based on whether they were greater than 0.5 (orange) or less or equal than 0.5 (purple). This coloring was then used for the SNPs in the segment in all other regions. Segments that did not have allelic imbalance in any of the regions could not be phased and are therefore colored in gray. Segments where all samples have allelic imbalance with the same allele gained or lost are shown in faded colors. Conversely, segments where either some of the samples show no imbalance while other samples show allelic imbalance, or mirrored subclonal allelic imbalance is observed are shown in darker colors. Segments where in at least one region SNPs colored in purple have BAF values greater than 0.5 can be classified as having mirrored subclonal allelic imbalance. Examples of this can be seen on chromosome 4q and a small segment on 18p.

### 6.6.6 Patient PANVH2

Patient PANVH2 did not have a metastasis sample sequenced, however, the PDX derived from the metastasis biopsy was sequenced. Subsequently, three cfDNA samples were collected and sequenced. The first cfDNA sample was collected one month after the patient received their last dose of Folfirinox and the CT scan showed partial response of the tumor. Shortly after the second sample was collected, the tumor progressed. The final cfDNA sample was collected 2.5 months later and simultaneously the CT scan showed stable disease (Figure 6-28).

![Graph showing ctDNA positivity and response to treatment over time for Patient PANVH2](image-url)
Estimating copy number profiles by using ASCAT showed purity below 5% for all cfDNA samples and no copy number aberrations could be detected. The PDX sample had high tumor content and clear copy number aberrations were found by ASCAT (Figure 6-29).

By applying the phasing algorithm explained in Section 6.6.2 copy number aberrations tracking with LOH events found in the PDX could be found in the second cfDNA sample (06/06/2018). Overall, 43 segments in the cfDNA sample were found with copy number aberrations, including two segments with MSAI, one on chromosome 2q and one on 7p (Figure 6-30). Additionally, two segments with copy number aberrations were found in the first cfDNA sample from April, one on chromosome 1p and one on chromosome 17p, both of which were also found in the second cfDNA and the loss on chromosome 17 encompassed TP53. No copy number aberrations were found in the final cfDNA sample which was consistent with no mutations found in that sample.
In this chapter, a cohort of pancreatic ductal adenocarcinoma patients was introduced, and the genomic characteristics of the metastatic tumors were explored. Of particular interest were differences in patients presenting with germline BRCA variants compared to patients whose tumors had only somatic BRCA alterations as well as those with wildtype BRCA. Within the cohort, no significant difference was found, but a trend towards a higher mutation burden in GL BRCA tumors compared to somatic and WT BRCA tumors could be observed. This result was validated in the TCGA cohort where GL BRCA tumors had a significantly higher mutation burden compared to both somatic and WT BRCA tumors. Similarly, a trend towards higher activity of SBS3, which is linked to defective homologous recombination DNA damage repair, was observed in the cohort, and subsequently validated in the TCGA cohort. No significant difference between somatic and WT BRCA tumors was observed, in either cohorts, relating to mutational burden or SBS3 activity. Similarly, no clear differences were observed at the copy number level between patients with GL, somatic or WT BRCA. Although the proportion of LOH in somatic BRCA tumors was higher than in GL or WT BRCA tumors, it is unclear whether this is due to correlation or causation. The majority of somatic BRCA alterations...
in the cohort were LOH of \textit{BRCA} rather than somatic mutations, therefore the higher proportion of LOH could either be driven by the LOH of \textit{BRCA}, or tumors with higher proportions of overall LOH are more likely to also have LOH at the \textit{BRCA} locus.

To date, olaparib has been approved as maintenance therapy for patients with metastatic pancreatic cancer and germline \textit{BRCA} mutations (Brown and Reiss, 2021). Although several PARP inhibitors have been approved for treatment of patients with either germline or somatic \textit{BRCA} mutations in breast cancer, in pancreatic cancer the only phase 3 trial to date has considered patients with germline rather than somatic \textit{BRCA} mutations (Patel et al., 2019; Brown and Reiss, 2021; Chi et al., 2021). Therefore, the benefit of PARP inhibitor therapies for tumors with somatic \textit{BRCA} alterations remains unclear. However, the fact that no significant differences were observed between somatic and WT \textit{BRCA} tumors could mean that PARP inhibitor therapies might not provide a significant benefit for patients with somatic \textit{BRCA} alterations, or this might be limited to mutations rather than LOH of \textit{BRCA}.

To date, most studies have detected ctDNA from cfDNA by using \textit{KRAS} mutations as well as other frequently mutated genes in pancreatic cancer (Le Calvez-Kelm et al., 2016; Adamo et al., 2017; Cohen et al., 2017; Pietrasz et al., 2017; Pishvaian et al., 2018; Mohan et al., 2019). However, the majority of these studies have investigated the use of ctDNA as a prognostic biomarker rather than utilized the samples to understand tumor evolution. In this chapter, the cfDNA samples in the cohort were used to explore heterogeneity and ongoing evolution in late-stage disease. For this, novel methods were developed which leverage a high purity tumor sample to rescue both mutations and copy number events in low purity cfDNA samples. For one patient in the cohort, significant heterogeneity was observed for copy number alterations in the metastasis and the simultaneously sampled cfDNA. Furthermore, during and after treatment additional heterogeneity was observed across the cfDNA samples. In another patient, the described method could be used to rescue somatic copy number aberrations in a cfDNA sample and highlighted that significant allelic imbalance could be observed even at low tumor purities.

One main limitation of this approach is the need for at least one high purity sample to rescue lower purity cfDNA samples. Furthermore, by using whole exome sequencing with around 100x coverage, the approach is limited to cfDNA samples with relatively high
tumor purity and likely underestimates the extent of heterogeneity within the samples. Therefore, to further explore heterogeneity and ongoing evolution, deeper sequencing of samples would be needed.
Chapter 7 Discussion

Applying evolutionary measures to predict metastatic risk as well as the metastatic potential of individual populations within the primary tumor has not been explored extensively. Understanding which patients are likely to develop disease recurrence could inform patient stratification and by predicting which tumor cells harbor metastatic potential new targets for therapy could be explored. Although stage is used to predict outcome in most cancer types, some patients with early-stage tumors develop disease recurrence, while other patients with metastatic disease at diagnosis do not develop further progression of their disease. Investigating paired primary-metastatic samples and tracking tumor evolution over time can give further insights into the metastatic process and inform future therapies.

This thesis has investigated the evolutionary processes leading to metastasis by characterizing the dissemination patterns from the primary tumor and timing of metastatic divergence. Furthermore, migrations between different metastatic sites, as well as ongoing evolution in late-stage disease were explored.

7.1 Inferring dissemination patterns from the primary tumor and investigating the seeding clones

7.1.1 Novel clustering and tree building approach in large-scale multi-region patients

To characterize which cells within the primary tumor harbor metastatic potential and can lead to disease recurrence within bulk sequencing data, first the cancer cell populations need to be deconvolved and their phylogenetic relationships inferred. In this thesis, a novel tool to cluster mutations and infer phylogenetic trees was presented (Chapter 3) to aid in answering these questions. In comparison to previously described methods (Roth et al., 2014; Deshwar et al., 2015; Malikic et al., 2015; Popic et al., 2015), the tool presented here can be applied to tumors with a large number of regions sampled, including both the primary tumor and any metastatic samples. By clustering all samples from a given patient simultaneously, increased resolution can be achieved, and the multi-region data leveraged in detecting mutations present across samples. Furthermore, mutations are first clustered based on presence or absence across samples which
increases computational speed and reduces artificial heterogeneity by removing low frequency clusters driven by a small subset of mutations that may represent sequencing artefacts. The tool has been benchmarked using a simulation framework incorporating metrics from the TRACERx cohort, such as mutation burden, number of samples and tumor purity.

To date this clustering and tree building method has only been applied to whole exome sequencing data. Future work will look into adapting the method to also be applicable to whole genome sequencing. Although the method should be transferable, computational speed will likely need to be improved to account for the much larger number of mutations called within whole genome sequencing data, especially when the number of tumor samples considered is low, as not much parallelization is possible in that case. Furthermore, estimation of clone proportion could be integrated into the tree building method and combined with the multiple tree solutions to summarize different solutions leading to similar clone proportions. Finally, although the tree building method provides multiple tree solutions, this information has not been harnessed fully. Future work could look into integrating additional tree resolution and providing confidence estimates for each branch and node based on how frequently it appears across the solutions. Additionally, the multiple tree solutions could be summarized into a few consensus trees that fully explain the solution space in a similar way as has been described previously in (Aguse, Qi and El-Kebir, 2019).

7.1.2 Dissemination patterns from the primary tumor

To infer the dissemination patterns from the primary tumor the output from the clustering and tree building was used, as described in Chapter 4. In 66% of cases, the dissemination pattern was defined as monoclonal, meaning metastatic potential was likely acquired once during tumor evolution. This is in keeping with previous studies which have found predominantly monoclonal dissemination (Brastianos et al., 2015; Turajlic, Xu, Litchfield, Rowan, Chambers, et al., 2018; Hu et al., 2020; Lee et al., 2020). Monoclonal dissemination likely means that metastatic potential was acquired once during the life history of the tumor and that any actionable drivers from the primary tumor are shared across all metastatic sites. Conversely, for polyclonal dissemination, metastatic potential was either acquired early in tumor evolution and shared across all cells in the primary tumor, or was acquired multiple times which could indicate that combination therapies might be necessary to target the different cell populations giving
rise to different metastases. Of the cases with polyclonal dissemination, a subset showed evidence of polyphyletic dissemination, meaning multiple clones from different branches of the phylogenetic tree seeded metastases.

When comparing dissemination patterns between lymph node metastases resected at surgery (primary LN metastases) and subsequent recurrence/progression samples, a higher proportion of primary LN metastases were defined as polyclonal in comparison to recurrence/progression samples, consistent with previous work in colorectal cancer (Reiter et al., 2020). This could indicate that recurrence/progression samples evolve through a stronger bottleneck, or that a higher fitness of cells is necessary to seed recurrence/progression samples in comparison to primary LN which are primary tumor adjacent. Indeed, modelling suggests that in some cases although multiple cell populations disseminated from the primary tumor, the seeding pattern ultimately detected within metastasis can appear monoclonal (TRACERx consortium, 2022b). This is particularly marked in simulations with higher selection coefficients.

Finally, sampling bias was shown to play an important role in inferring dissemination patterns. The number of sampled metastatic regions was significantly lower in monoclonal compared to polyclonal cases highlighting that sampling individual sites when the patient presents with widespread disease likely underestimates the amount of heterogeneity. However, due to patient frailty or difficulty in accessing the metastatic site, in most cases only a single biopsy could be obtained. This highlights the need for alternative sampling methods, such as ctDNA (Abbosh et al., 2017; Parikh et al., 2019), or extensive studies such as PEACE to better understand the full extent of heterogeneity at disease recurrence and progression.

Additional work implementing models of metastatic seeding could help elucidate potential discrepancies between the cells disseminating from the primary tumor and those finally detected upon sampling of metastasis. Initial modelling performed suggests that multiple cell populations could seed a metastatic site, but due to a higher fitness advantage of one population, a clonal sweep at the metastatic site could occur leading to subsequent characterization of monoclonal dissemination. Similarly, it could be assumed that due to genetic drift a small number of cells could lead to an observed polyclonal dissemination pattern.
Finally, patient-derived xenografts could also provide insight into the process of metastatic dissemination. Unlike with metastatic dissemination that occurs at an unknown time and therefore the clonal composition of the primary tumor remains unknown and can only be estimated, PDX models are derived at the point of tumor sampling. Therefore, the clonal composition of the tumor is known and subsequent bottlenecking within the PDX sample can be correctly characterized.

### 7.1.3 The seeding clones present selection compared to non-seeding clones

Previous large-scale studies investigating differences between the primary tumor and metastases have included single-sample data per patient, either a sample from the primary tumor or a single metastatic sample (Priestley et al., 2019; Nguyen et al., 2022). By using unpaired data, it is not possible to infer which events were truly metastasis-unique and which were already present in the primary tumor and selected in metastases. This highlights the power of the TRACERx paired primary metastasis cohort as this allowed for classification of each clone as shared, primary- or metastasis-unique. Consequently, it was possible to compare mutations occurring in the seeding clone, which were shared between the primary tumor and metastasis, with mutations that are primary-unique and therefore did not harbor metastatic potential.

Chapter 4 showed that mutations within the seeding clone had a significantly higher cancer cell fraction and were present in a larger proportion of primary tumor regions compared to their non-seeding counterparts. This highlighted that the seeding clone conceivably exhibited a fitness advantage within the primary tumor, which was further supported by a dN/dS analysis of cancer genes (Martincorena et al., 2017). This showed that mutations in seeding clones were under positive selection compared to mutations in primary-unique clones, particularly in lung squamous cell carcinoma. Strikingly, no significant selection was observed in lung squamous cell carcinoma when considering all subclonal mutations (TRACERx consortium, 2022a).
7.2 Characterizing the metastatic process and metastasis-to-metastasis seeding

7.2.1 Timing of metastatic divergence

As mentioned above, stage is one of the main predictors of outcome for multiple cancer types. Patients diagnosed with later stage disease often have a higher risk of metastasizing, suggesting that metastatic potential might be acquired late in the evolution of the tumor. However, some patients with early-stage disease also develop recurrences, which highlights that in these cases the tumors might present a “born to be bad” phenotype (Sottoriva et al., 2015; Turajlic, Xu, Litchfield, Rowan, Chambers, et al., 2018). Therefore, understanding when metastases diverge could help stratify patients for treatment. In some cases, patients with early-stage disease might benefit from adjuvant treatment, while in other cases for patients presenting with later-stage disease, additional treatment might not be necessary.

In Chapter 5 definitions for timing metastatic divergence relative to the last clonal sweep within the primary tumor were presented. Consistent with previous studies in esophageal, pancreatic, and breast cancer, the majority of metastases diverged late, after the last clonal sweep in the primary tumor (Yachida et al., 2010; Yates et al., 2017; Noorani et al., 2020). In the case of late divergence, all primary-ubiquitous mutations were shared with the metastatic samples. In early divergence cases not all primary-ubiquitous mutations were detected within metastases, however it was shown that the majority of mutations were still shared between the primary tumor and metastasis. Furthermore, when considering mutations occurring pre-whole genome doubling within the primary tumor, these were almost always shared with metastatic samples suggesting that such mutations could represent good targets for targeted therapies, including immune-vaccine approaches.

Furthermore, it was shown that under-sampling of the primary tumor could lead to incorrect classification of early metastatic divergence. The definitions of timing divergence presented are dependent on correctly classifying primary-ubiquitous mutations. Therefore, sampling only a single tumor region could lead to over-calling of clonal mutations due to clonal illusion (de Bruin et al., 2014). Consequently, a higher number of primary tumor samples lead to higher resolution and correct inference of
primary-clonal mutations. Previous studies in colorectal and breast cancer have found predominantly early divergence, however these results were likely influenced by methods used to infer phylogenetic trees, or under-sampling of the primary tumor (Faltas et al., 2016; Brown et al., 2017; Hu et al., 2019). Strikingly, (Hu et al., 2019) found that 81% of tumors seeded metastases when the primary tumor was less than 0.01cm³ and therefore clinically undetectable. However, three out of four tumors estimated to disseminate late had multi-region sequencing of the primary tumor compared to only 6 out of 17 tumors estimated to diverge early. This again highlights the importance of extensive sampling of the primary tumor to correctly infer the timing of divergence.

Finally, as the data presented was obtained through whole exome sequencing, only relative timing of metastatic divergence was possible in most cases. In one example the mutational signature arising from platinum chemotherapy could be used to infer the timing of divergence of two brain metastases (Pich et al., 2019, 2021; Landau et al., 2020). However, in order to apply this method at a larger scale, with high confidence, whole genome sequencing data would be necessary. Future studies using whole genome sequencing could use the mutational signatures associated with aging (Alexandrov et al., 2020) to infer absolute timing of metastatic divergence.

7.2.2 The PEACE cohort allows for better characterization of the metastatic process

The TRACERx paired primary-metastatic cohort gave insight into the metastatic process, however due to low numbers of metastatic samples per primary tumor, the full extent of heterogeneity was likely underestimated. To better understand the metastatic process and metastatic heterogeneity, data from the PEACE study was processed and analyzed. PEACE allows for extensive sampling across all metastatic sites which can be analyzed to recreate a more complete picture of the metastatic process. Some of the patients recruited into PEACE were previously also recruited into TRACERx which allowed for analysis of at least two to three timepoints across the cohort. An overview of the multi-region, multi-timepoint paired primary-metastatic cohort was shown in Chapter 5. These data can be used to characterize patterns associated with specific sites of disease, and heterogeneity within as well as across metastases within the same tissue or throughout the body. Future work includes understanding the interplay between genomic and transcriptomic data of metastases with the immune microenvironment, as well as integrating metastatic migrations.
7.2.3 Metastases can seed further sites of disease

In Chapter 4 only dissemination from the primary tumor was characterized. However, it has been previously shown that in some cases metastases can go on to seed other metastatic sites (Gundem et al., 2015; Brown et al., 2017; Naxerova et al., 2017; Chen et al., 2022). To further explore this in Chapter 5, MACHINA, a tool to infer migration histories was applied to both TRACERx and PEACE (El-Kebir, Satas and Raphael, 2018). Within the TRACERx cohort several patients had multiple metastases for which the analysis could be performed. Of particular interest here were patients with both primary LN disease and subsequent recurrence/progression samples. Patients with primary LN disease often received adjuvant radiotherapy, however the survival benefits associated with radiotherapy remain ambiguous (Waddle et al., 2018), likely due to the toxicity of the treatment. If patients with LN involvement could be stratified into two groups, one where the LN represents an “evolutionary dead-end” and one where the LN could seed further metastases, radiotherapy could be administered in a more selective manner. Within the TRACERx cohort, primary LN metastases were found to seed subsequent recurrence/progression metastases in less than 20% of cases suggesting that the majority of primary LN metastases represent metastatic potential of the primary tumor rather than a gateway to further metastases.

A similar approach was used to infer migration patterns within PEACE where larger-scale patterns of site-specific migrations could be observed. Although the majority of metastases were seeded directly from the primary tumor, 49% of metastases were seeded by other sites, such as from lymph node and lung metastases. Furthermore, recurrent patterns of source-to-target sites could be observed, for example when a liver metastasis seeded subsequent metastases these were predominantly also found in the liver while lymph node metastases could seed metastases across multiple organs.

Finally, using the inference of metastatic migration patterns and linking these to imaging reports, absolute timing of divergence of certain metastases could be inferred. However, to further time metastatic divergence, whole genome sequencing would be required. As the PEACE study recruits patients with metastatic disease and most metastatic samples are obtained at tissue harvest, the cohort presented here is clearly biased with regards to metastatic dissemination as only patient’s tumors are included that ultimately succumb to their disease.
Future work to explore migrations could integrate the single cell data sequenced as a part of PEACE. Single cell data can give an exact snapshot of the clones present at the point of sampling rather than relying on the inferred phylogenetic estimates using bulk whole exome sequencing. Therefore, similarities of clones present across various metastatic sites can help resolve the observed migration patterns using MACHINA (El-Kebir, Satas and Raphael, 2018). Similarly, MACHINA could be adapted to enumerate the full solution space of migration patterns. This space could then be summarized to provide resolution of migration graphs as well as confidence estimates for each migration.

### 7.3 Ongoing evolution and heterogeneity in late-stage disease

To date, it has been assumed that pancreatic cancer is predominantly driven by a few commonly altered genes which result in stepwise increase of cell cycle proliferation and are often shared with pancreatic precursor lesions (Jones et al., 2008; Kanda et al., 2012; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). However, Chapter 6 highlights heterogeneity across patients during tumor initiation, but also heterogeneity within patients is shown, including ongoing SCNA heterogeneity throughout the further disease course after the initial metastasis was sampled.

Several patients within the cohort introduced in Chapter 6 presented with a germline mutation in one of the BRCA genes. Most of the patient’s tumors did not exhibit a secondary somatic BRCA mutation, therefore indicating that for pancreatic ductal adenocarcinoma, germline BRCA mutations might not be responsible for tumor initiation. However, genomic differences were observed between patients with germline BRCA mutations and those with wildtype BRCA, including higher mutation burden as well as a higher activity of a mutational process linked with homologous recombination deficiency. This highlights differences in the disease course of these two groups.

Analysis of cfDNA samples obtained for each patient within the cohort elucidates ongoing evolution and within-patient heterogeneity at the metastatic stage. Firstly, high-confidence patient specific mutations could be tracked within the cfDNA samples to determine ctDNA positivity. Additionally, an algorithm was developed to detect somatic copy number aberrations in low purity samples, leveraging high purity samples from the same patient. This was highlighted for one patient where using this method, copy number
aberrations could be found within one of the cfDNA samples where copy number inference using ASCAT was not possible due to low tumor purity. Furthermore, in one patient, this approach was used to highlight both heterogeneity between the metastatic sample and a cfDNA sample obtained on the same day, as well as heterogeneity between these initial samples and additional cfDNA samples obtained at later dates.

As tumor content in the blood is often very low, using whole exome sequencing restricts the number of mutations and SCNAs that can be detected, especially subclonal variants. Although using a “personalized” approach to detect mutations increases the sensitivity, deep targeted sequencing of tumor-specific mutations can lead to better resolution, as has been shown in (Abbosh et al., 2017). However, this limits detection of emerging variants during late-stage disease which could for example be observed in one patient within the cohort.

7.4 Conclusions

The results presented in this thesis give insight into the phylogenetic relationships observed between primary tumors and metastases. By understanding which cancer cells are shared between the primary tumor and any metastatic samples, a better characterization of the metastatic process is possible. Indeed, the data in this thesis show that cells within the primary tumor which seed metastases are generally more prevalent compared to cells which do not seed metastases and their mutations show evidence of significant selection of cancer genes, possibly indicating a fitness advantage that already persists in the primary tumor and leads to metastatic dissemination.

Furthermore, the results of this thesis show that the majority of metastases diverge late, after the last clonal sweep which means that the majority of primary-ubiquitous mutations are shared with metastases and therefore represent ideal targets for therapy. In the cases where metastases diverge early, this often occurs below the limit of detection, highlighting that in these cases early-detection is likely important to increase the prognosis of patients. Additionally, metastasis-to-metastasis seeding is observed in a significant subset of cases analyzed, emphasizing that in the metastatic setting, treating solely the primary tumor could potentially not lead to significant reduction of tumor burden.
Finally, significant genomic heterogeneity is shown in the metastatic setting and during subsequent therapy. The data presented in this thesis also highlight the benefits of using cfDNA as a minimally invasive approach to further elucidating tumor evolution and ongoing heterogeneity, especially when paired with previously collected tumor samples.

Taken together, these results indicate that the metastatic process is highly diverse and that more work is needed to better characterize the routes to metastases. Furthermore, no single treatment strategy will benefit all patients and therefore better ways to stratify patients based on genomic, transcriptomic, or epigenomic data will be necessary to improve prognosis.
References


Jiao, Y. et al. (2011) “DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors,” *Science (New York, N.Y.).* American Association for the Advancement of Science (AAAS), 331(6021), pp. 1199–1203.


References


References


References


Zámborszky, J. et al. (2017) “Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions,” Oncogene, 36(35), pp. 5085–5086.


