Integrating multi-omics data by mapping subclonal events on tumour evolutionary trees

Hoang Son Tran

A thesis submitted for the degree of
Doctor of Philosophy

University College London

PhD Supervisors:

Prof Javier Herrero
Dr Nicholas McGranahan
Prof Christopher Barnes

August 2023

DECLARATION

I, Hoang Son Tran, 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.

Abstract

Inferring the tumour's evolutionary history is crucial for unravelling the intricate landscape of intratumour heterogeneity underlying cancer progression. Several bioinformatics tools have been designed for deciphering the subclonal population of the heterogeneous tumour mass. However, most of them rely on single-omics analysis and methods for integrating multi-omics data in the context of tumour evolutionary trees are still lacking.

In this thesis, the development of MAPping SubClonal Events (MAPSCE), a new tool for mapping of subclonal events on tumour evolutionary trees, is described. This method allows for integration of multi-omics data in multi-sample cancer evolutionary studies. In essence, MAPSCE implements a branch test where quadratic programming is applied to every branch of a patient tumour tree to find the best mapping branch (including the root). Each solution translates into a Bayesian Information Criterion value, and Bayes factors for model selection. MAPSCE has been released as an R package.

Multiple datasets with different types of copy number events and varying degrees of noise up to ±30% were simulated to assess the reliability of the tool. For losses of haploid genes, MAPSCE was benchmarked against a tool of similar functionality, LOHHLA, showing both an increase in specificity and sensitivity. This comparison was not possible for other types of copy number events as MAPSCE is the only tool to date with the ability to map these.

Lastly, MAPSCE's potential applications were demonstrated in several analyses of multi-region, multi-omics datasets. Subclonal biallelic inactivation of tumour suppressor genes on subclonal level was identified in lung cancer patients. Subclonal changes of gene expression were further compared against subclonal copy number events to infer cases of copy number dependent or independent allele specific expression.

This work provides an innovative way to integrate multi-omics data in multi-sample cancer studies, refining the study of evolutionary processes underlying intratumour heterogeneity.

Impact Statement

Cancer is the second leading cause of mortality around the world, with lung cancer being one of the most common types of cancer among men and women. Intratumour heterogeneity, the diversity of cells within a single tumour mass, is one of the main issues preventing development of effective cancer therapeutics, as it leads to failure to validate cancer biomarkers, decreases drug efficacy, exacerbates patient prognoses and increases therapy costs.

Many methods have been developed for studying the diversity within each individual tumour, however most of them only focus on one type of alteration at a time. This single-lens approach disregards the broader perspective of how multiple changes contribute to tumour evolution. Furthermore, many of the cancer studies have focused on inferring the tumour evolution based on a single biopsy, where two samples from different sites of a tumour could provide vastly different pictures of the tumour's entire evolutionary process. While the rapid generation of new cancer datasets provides more avenues for studying tumour evolution, tools for integration of the various types of cancer data based on multiple biopsies in an evolutionary context are still lacking.

This thesis presents the development, and testing of a novel methodology, MAPping Subclonal Events, for studying intratumour heterogeneity. This integrative tool allows for combining the different layers of information across multiple tumour samples as well as tracking of the diverse changes, to provide a more comprehensive understanding of the hidden processes underlying tumour evolution. The analysis of cancer datasets demonstrates the evolutionary questions that could be addressed with the tool presented in this thesis. Exploring the opportunities opened up by this research could identify novel drivers of cancer that could be targeted to predict therapy response.

Ultimately, the research presented in this thesis provides a novel, integrative approach to studying tumour evolution, paving the way for more effective cancer diagnostics and therapeutics tailored to individual patients.

Acknowledgements

Firstly, I am extremely grateful to my primary supervisor, Prof Javier Herrero, for providing me with the opportunity to undertake a PhD at the Bill Lyons Informatics Centre. Thank you, Javier, for mentoring me throughout my studies. This thesis would not have been possible without your incredible commitment and kind encouragement. I would also like to acknowledge my secondary and tertiary supervisors, Dr Nicholas McGranahan and Prof Christopher Barnes, for providing me with their constant support and valuable feedback during the past four years.

During my PhD I have had the pleasure of sharing the office with all the amazing people at the Bill Lyons Informatics Centre. I wanted to personally thank Chuling for her constant willingness to help with any questions and struggles throughout our PhD projects, Lucia for her honest advice with anything science and life related, and Stephen for sharing his bioinformatics knowledge. Thank you all for making my time at the BLIC such a welcoming and enjoyable experience.

I would like to especially thank my family for their unwavering support. Thank you to my Mom for always being there for me every step of the way, to my Dad for inspiring a lifelong commitment to learning from an early age, and to my Brother for supporting me in every way he could.

Finally, I wanted to extend my gratitude to my closest friends for their moral support over the years. I would like to especially thank Sosna and Dominika for all the memorable experiences we have had while living together in London. I am truly thankful for the experience of sharing our PhD journeys together, and it has been such a unique and rewarding time. Most importantly, thank you, Dominika. You have always supported me on this journey. Your love and patience have helped me through it all, and I am forever grateful for it.

Table of Contents

Table of Figures

List of Tables

Abbreviations

ABC – approximate Bayesian computation

ANOVA – one-way analysis of variance

AP – antigen presentation

APM – antigen presentation machinery

APOBEC – apolipoprotein B mRNA editing enzyme catalytic polypeptide-like

ASE – allele-specific expression

BAC – bacterial artificial chromosome

BAF – B-allele frequency

bp – base pairs

BF – Bayes Factors

BIC – Bayesian information criterion

CCF – cancer cell fraction

CN – copy number

CNA – copy number alteration

$\mathrm{CN}_{\mathrm{after}}$ – copy number state after the branch

$\mathrm{CN}_{\mathrm{before}}$ – copy number state before the branch

CRAN – the comprehensive R archive network

DNA – deoxyribonucleic acid

ecDNA – extrachromosomal deoxyribonucleic acid

FISH – fluorescence in situ hybridisation

GO – gene ontology

HGP – human genome project

HGSC – high-grade serous ovarian cancer

HLA – human leukocyte antigen

ICGC – the international cancer genome consortium

KEGG – Kyoto encyclopaedia of genes and genomes

LogR – log-ratio

LOH – loss of heterozygosity

LOHHLA – loss of heterozygosity in human leukocyte antigen

LOHHLA/QP – loss of heterozygosity of human leukocyte antigen's mapping approach

LUAD – lung adenocarcinoma

LUSC – lung squamous carcinoma

MAPSCE – mapping subclonal events

MCMC – Markov chain Monte Carlo

MHC – major histocompatibility complex

MLE – maximum likelihood estimation

MRCA – most recent common ancestor

NGS – next-generation sequencing

nnls – non-negative least squares

NSCLC – non-small cell lung cancer

PCAWG – the pan-cancer analysis of whole genomes

QP – quadratic programming

RNA – ribonucleic acid

RSS – residual sum of squares

SBS – single base substitution

sc – single-cell

sc-seq – single-cell sequencing

SCNA – somatic copy number aberration

SMS – single-molecule sequencing

SNP – single nucleotide polymorphism

SNV – single nucleotide variation

TCGA – the cancer genome atlas

TIL – tumour-infiltrating lymphocytes

TRACERx – tracking cancer evolution through therapy

TSG – tumour suppressor gene

VAF – variant allele frequency

WES – whole-exome sequencing

Chapter 1 Introduction

1.1 Tumour evolution

Tumours comprise individual cancer cells with distinct genetic alterations. The evolution of a tumour is based on these individual cancer cells acquiring genetic changes over time, some of which confer a growth advantage. These so-called driver events lead to clonal expansion, where cells proliferate and can, in turn, establish new subpopulations of cells with unique genotypes. One key manifestation of tumour evolution is the presence of intratumour heterogeneity (ITH), whereby even within the boundaries of a single tumour, there is considerable genetic diversity among the various subpopulations of cells (Mullighan et al. 2008; Navin et al. 2011; Gerlinger et al. 2012; Shah et al. 2012). This phenomenon highlights the complex and dynamic nature of tumour evolution.

Intratumour heterogeneity poses a significant and unmet challenge in the treatment of cancer, as it results in increased therapy costs, reduced drug efficacy, failure to validate cancer biomarkers and poorer prognoses for cancer patients (Figure 1-1). Jamal-Hanjani et al. (2017) demonstrated the negative impact of subclonal copy number alterations on patient outcomes. In another study, only one out of 28 tested biomarkers effectively predicted patient survival, with others failing primarily due to ITH (Gulati et al. 2014). Biswas et al. (2019) highlighted the negative impact of the tumour sampling bias resulting from ITH on the efficacy of cancer biomarkers and presented their prognostic signature in lung cancer. Another study suggested the role of genetic and transcriptomic diversity as the origin of chemotherapy resistance in pancreatic cancer (Seth et al. 2019). Finally, Marusyk, Janiszewska, and Polyak (2020) further discussed the increasing therapy costs associated with the necessary routine ITH assessment for patient prognostication. Taken together, understanding the complex molecular landscape of cancer necessitates the development of novel and tailored approaches to correctly consider ITH.

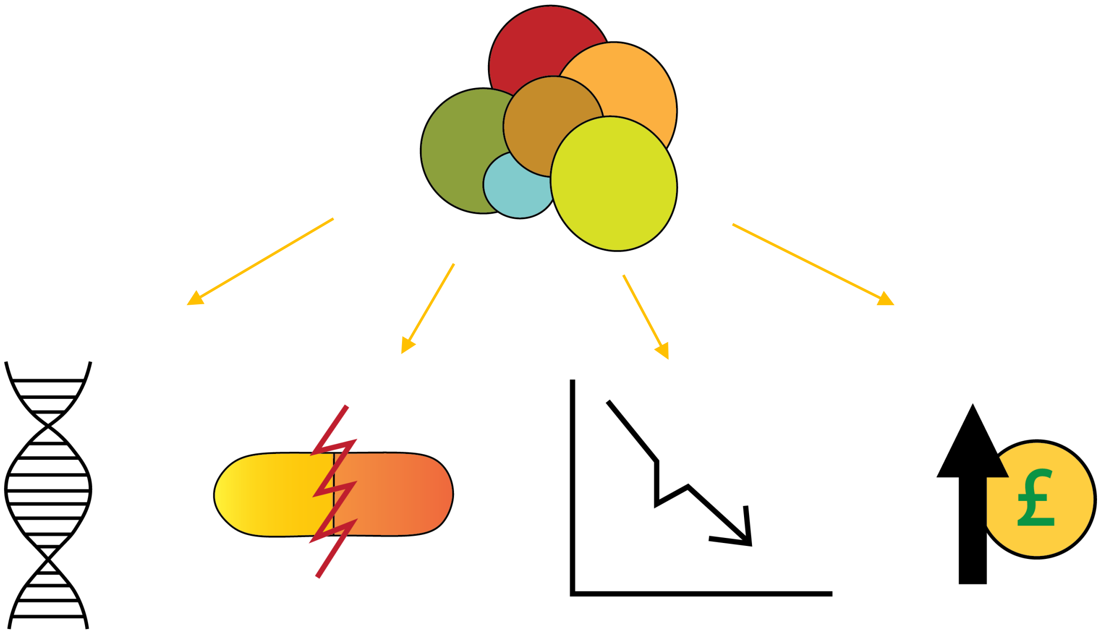


Figure 1-1 Intratumour heterogeneity leads to failure to validate cancer biomarkers, decreasing drug efficacy, poor prognosis of cancer patients and increasing therapy costs

1.2 Models of tumour evolution

One important discussion regarding tumour evolution revolves around whether cancers develop under the clonal expansion (positive selection) or the stochastic model (neutral selection) (Turajlic et al. 2019) .

1.2.1 Clonal expansion model

In 1859 Charles Darwin introduced the theory of branching evolution, in which diverse populations originated from a common ancestor through the process of natural selection. Analogically, Peter Nowell in 1976 hypothesised that cancer follows the Darwinian selection process, in which genetic variability is at the core of tumour evolution and different tumour cell populations compete for dominance by acquiring selective growth advantages over others (Nowell 1976). The clonal evolution of cancer (Figure 1-2) follows a Darwinian selection process, where the majority of the cancer cell population can share one or two progenitors, however different tumour clones acquire varying mutations, forming divergent tumour cell populations over time (Polyak, Haviv, and Campbell 2009). Events that give a selective advantage (driver events) lead to further development of the dominant subclones, while disadvantageous mutations and healthy cells slowly become evolutionary dead-ends (Marusyk and Polyak 2010). This results in a highly heterogeneous tumour, which requires a multi-faceted approach specifically tailored to kill every individual dominant subclone. Surviving clones could proliferate and expand, initiating another clonal expansion of resistant cells (Greaves and Maley 2012). The continuous clonal selection theory has been evidenced by the low ITH of the clonal driver events (Gao et al. 2016; Notta et al. 2016).

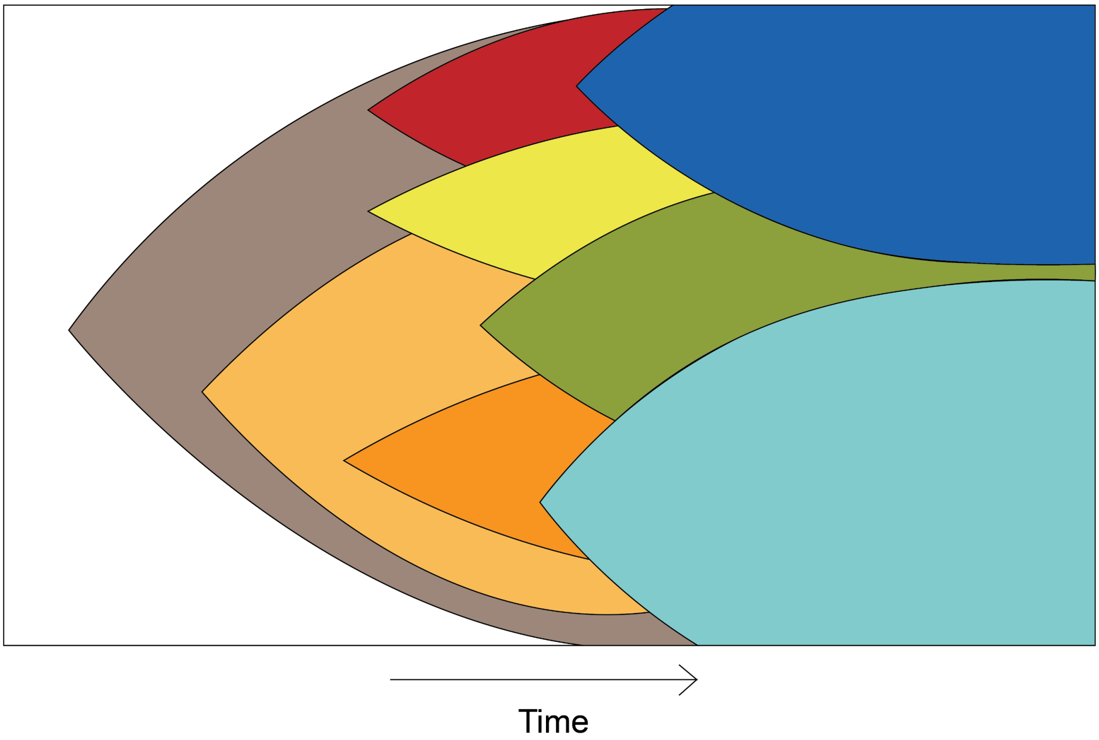


Figure 1-2 Clonal expansion model, in which one cell initiates tumour progression that is driven by the emergence of new subclonal populations in cancer.

1.2.2 Stochastic model

Conversely, the stochastic evolution model suggests that tumours evolve by acquiring mutations via genetic drift due to the random changes in allele frequencies with no single dominant subclone. According to this model, multiple cell populations with different genetic alterations coexist. Consequently, it is suggested that ITH is driven by the diverse random genetic and epigenetic alterations acquired under neutral evolution in subclonal populations that ultimately do not lead to a clonal sweep (Yates et al. 2015; Gerstung et al. 2020).

1.3 Types of events driving tumour evolution

There are different types of alterations which can drive tumour progression and contribute to intratumour heterogeneity. These include genetic alterations such as point mutations, insertions, deletions and copy number alterations (CNA), and epigenetic alterations, such as modifications to the chromatin structure, gene expression, and methylation changes without altering the DNA sequence (Takeshima and Ushijima 2019). Point mutations are the changes in the DNA sequence which substitute, insert or delete a single nucleotide. Copy number alterations encompass changes to the number of copies, either gains or losses, of specific segments of the DNA. The gene expression changes refer to the alterations affecting the production of RNA, leading to a modified expression level of a particular gene. Methylation changes are the alterations in the methylation patterns of the DNA, involving the addition or subtraction of a methyl group. Both the genetic and epigenetic alterations can drive tumour progression when they confer a growth advantage, either by activating, duplicating or upregulating oncogenes such as *KRAS*, or inactivating, losing or downregulating tumour suppressor genes such as *TP53* (Hanahan and Weinberg 2000; 2011; Jamal-Hanjani et al. 2017; Juul et al. 2021; Frankell et al. 2023). While there are numerous other types of alterations which can contribute to tumour progression, this thesis will primarily centre on the major classes of alterations mentioned above.

All the aforementioned events can be further categorised as clonal or subclonal events. Clonal events are alterations present in all cells within a tumour sample, while subclonal events are alterations that occur only in a subset of cells within a particular tumour lineage (Black and McGranahan 2021). Knowing the clonality of the events is crucial for determining the timing and order of the alterations. This allows us to explore the evolutionary processes including parallel evolution, co-occurrence and mutual exclusivity. Co-occurring driver events collaborate to activate oncogenic pathways, probably providing an additional selective advantage. In contrast, mutual exclusivity reveals the intricate interactions between specific drivers and can potentially offer targets for cancer treatments (Tekle et al. 2021). Turajlic et al. (2018) demonstrated the parallel evolution of single nucleotide variants (SNVs) affecting *SETD2*, *BAP1* and *PTEN* driver genes. Deciphering these evolutionary processes and constraints uncovers the epistatic relationships between various driver events underlying branched tumour evolution (Landau et al. 2013; McGranahan et al. 2015; Cheng et al. 2022).

1.4 Brief history of sequencing techniques

The continuous development of new sequencing technologies has played a pivotal role in generating large amounts of data, and in turn, improving our understanding of the complex nature of tumour heterogeneity.

1.4.1 First-generation sequencing

The first-generation sequencing, Sanger sequencing, became the gold standard of sequencing technologies for years after its introduction in the 1970s and was later utilised for the ambitious Human Genome Project (HGP), which aimed to sequence the entire human genetic sequence (Sanger, Nicklen, and Coulson 1977; Olson 1993; Collins and Fink 1995). However, Sanger sequencing was a complex, labour-intensive process which involved multiple steps, including DNA extraction, purification, amplification, sequencing, gel electrophoresis, and manual data analysis (Crossley et al. 2020). Scaling the technique to large genomes required the preparation of BAC libraries which was also a multi-step process (Osoegawa et al. 2001). Consequently, Sanger sequencing was ill-suited for large-scale sequencing of complex tumour genomes.

1.4.2 Second-generation sequencing

The emergence of second-generation sequencing, also known as Next-Generation Sequencing (NGS), dominated by Solexa/Illumina, allowed for massively parallel sequencing, providing increased sequencing output at a reduced cost and time (Voelkerding, Dames, and Durtschi 2009). The NGS platform has been vastly improved since its inception by further reducing cost, increasing output, read length and depth of sequencing (Muir et al. 2016). These improvements to the NGS technology enabled sequencing of the genome of an individual or a tumour in a short time and at an affordable price.

1.4.3 Third-generation sequencing

Lastly, third-generation sequencing introduced single-molecule sequencing (SMS) (van Dijk et al. 2014) and nanopore sequencing (Clarke et al. 2009; Eisenstein 2012), both of which allowed for sequencing longer reads compared to NGS, while eliminating the need for DNA amplification (Xiao and Zhou 2020).

1.5 Large-scale cancer datasets

While the development of Sanger sequencing launched the HGP, other projects emerged over time to collect pan-cancer data and perform sequencing of the cancer genomes. The Sanger Institute’s Cancer Genome Project primarily utilised whole genome sequencing data to catalogue the cancer genes and patterns of clonal evolution in human tumours (Pleasance et al.

2010). The Cancer Genome Atlas (TCGA) aimed to assess the entire spectrum of genomic changes in human cancer (Tomczak, Czerwińska, and Wiznerowicz 2015). TCGA provides a public database on over 20,000 primary cancer and matched normal samples from 33 different cancer types. The International Cancer Genome Consortium (ICGC) was a global initiative to coordinate large-scale cancer genome studies and characterise the genomic landscape of over 50 different cancer types (Hudson et al. 2010). Lastly, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium was launched to identify common patterns of mutation in more than 2,600 cancer whole genomes, building on the work from ICGC and TCGA (Aaltonen et al. 2020).

1.6 Multi-sample sequencing

Importantly, the advancement of NGS techniques accelerated the sequencing progress at a lower cost, revolutionising cancer research. The main issue with inferring tumour evolutionary history from a single sample at one time point is intratumour heterogeneity. ITH makes it challenging to distinguish between the different subclonal populations within a heterogeneous tumour mass (Black and McGranahan 2021).

Longitudinal studies have emerged where samples from the same patient were collected at different time points including metastases to provide an overview of the changes occurring within the tumour over time. Cindy Yang et al. (2021) examined solid tumours of 73 patients across 30 different cancer types before and after pembrolizumab therapy. The study revealed that *BRCA2* mutations, increased mutation burden and elevated expression of immune signatures were associated with pembrolizumab sensitivity. In contrast, an abundance of CNA and loss of heterozygosity of *B2M* corresponded with drug resistance. Jiang et al. (2016) identified prognostic biomarkers by inferring tumour phylogeny from breast cancer patient xenografts and their subsequent propagation of metastatic xenografts. Boyle et al. (2021) examined samples from multiple time points of patients with smouldering multiple myeloma. The findings revealed an increased mutational load and elevated intratumour heterogeneity in patients who progressed from smouldering multiple myeloma to multiple myeloma.

Another approach involves sampling multiple regions of the same tumour at the same time. Gerlinger et al. (2012; 2014) have demonstrated how multi-region whole-exome sequencing (WES) on patients with renal cell carcinoma revealed branched tumour evolution within the heterogenous tumour mass with 63 – 69% of all somatic mutations not being detectable across every tumour region, thus being missed in the single-sample analysis. Other studies have since investigated intratumour heterogeneity using a multi-region sequencing approach. Yates et al. (2015) identified subclonal diversification of common breast cancer genes, including *PIK3CA, TP53, PTEN, BRCA2 and MYC* in breast cancer patients. In another study of 11 localised lung adenocarcinoma patients, Zhang et al. (2014) found that all three relapsed patients had significantly larger fractions of subclonal mutations in their primary tumours, compared to non-relapsed patients. Additionally, de Bruin et al. (2014) demonstrated how mutations associated with APOBEC-mediated mutagenesis (apolipoprotein B mRNA editing enzyme catalytic polypeptide-like) were mostly subclonal in four out of five lung adenocarcinoma (LUAD) patients.

1.6.1 TRAcking Cancer Evolution through therapy (TRACERx)

Recently, the TRAcking Cancer Evolution through therapy (TRACERx) showed novel insights into the study of intratumour heterogeneity of multiple cancers (Jamal-Hanjani et al. 2014). TRACERx is an ambitious translational research study which aims to elucidate the intricate evolutionary processes underlying cancer progression, and directly translate the findings to improved patient outcomes. Their work encompasses multiple cancer types, such as non-small cell lung cancer (NSCLC) (Jamal-Hanjani et al. 2017; Frankell et al. 2023), melanoma (Menger et al. 2016) and renal cancer (Mitchell et al. 2018; Turajlic et al. 2018). Their multi-sample research aims to track the tumour evolution through time and space to accurately identify clonal and subclonal drivers. The NSCLC TRACERx 100 study has demonstrated the significance of multi-region whole-exome sequencing. The study revealed that if only single-sample analysis had been used, 76% of subclonal mutations would have been misidentified as clonal (Jamal-Hanjani et al. 2017). Furthermore, significantly more mutations were identified with multi-region WES compared to single-sample analysis or with the use of a single NSCLC sampled from TCGA (Jamal-Hanjani et al. 2017).

This thesis will predominantly focus on the NSCLC TRACERx 100 dataset and one patient from the NSCLC TRACERx 421 dataset (Jamal-Hanjani et al. 2017; Frankell et al. 2023; Martínez-Ruiz et al. 2023).

1.7 Studying intratumour heterogeneity

Advanced cancer genomics allows for studying the extent of heterogeneity of the tumour and the effects of the clonality of mutations (Schmitt, Prindle, and Loeb 2012; Dan et al. 2015). There is no standardised and universal workflow for studying intratumour heterogeneity as it entails the analysis of different types of events, both genetic and epigenetic. However, the research presented in this thesis will focus predominantly on the integration of subclonal CNA in the context of a tumour tree built on SNVs. This section will focus on the workflow for that analysis.

1.7.1 Existing methods for studying ITH and their limitations

One of the first steps in the workflow is the detection of the SNVs and CNAs (Figure 1-3). The former includes variant calling, which is the identification and classification of the somatic point mutations. There are several tools available for the detection of SNVs, such as VarScan and VarScan2 (Koboldt et al. 2009; 2012), Strelka (Saunders et al. 2012) and MuTect (Cibulskis et al. 2013). Other tools focus on processing copy number data (Koboldt et al. 2012; Rimmer et al. 2014) and estimating the absolute copy number, enabling the detection of gain, loss, copy number neutral and loss of heterozygosity (LOH) events. Carter et al. (2012) developed ABSOLUTE, a method for absolute quantification of copy number, while jointly estimating tumour purity and ploidy. Van Loo et al. (2010) presented ASCAT, an allele-specific copy number estimation tool for solid tumours, which also estimates tumour ploidy and purity. Oesper, Mahmoody, and Raphael (2013) also described THetA, which identifies subclonal copy number alterations.

One way of studying ITH involves subclonal deconvolution, which is the process of disentangling the bulk sequencing data of a tumour to identify distinct features of subclonal populations (Barrett et al. 2017; Yang et al. 2019) (Figure 1-3). Subclonal deconvolution includes estimating the mutational cancer cell fraction (CCF), which involves calculating the proportion of cancer cells within a tumour sample carrying a specific mutation and clustering those mutations. Tools can temporally order the somatic point mutations to determine their clonality. Roth et al. (2014) developed PyClone, a Bayesian clustering method, which clusters somatic mutations to estimate subclonal frequencies while accounting for copy number changes and purity. Similarly, Miller et al. (2014) introduced SciClone, which also clusters somatic mutations into several subclones with estimated frequency, using a variational Bayesian mixture model. SciClone's variational Bayesian mixture model includes a computational termination condition, which is less computationally intensive than PyClone's Markov chain Monte Carlo (MCMC) convergence. However, SciClone's method is heuristic and can result in sub-optimal solutions compared to PyClone's stochastic MCMC. Ha et al. (2014) and Shen and Seshan (2016) both recognised the lack of tools for studying copy number heterogeneity in subclonal populations and developed TITAN and FACETS, respectively, to focus on inferring the copy number architecture. Fischer et al. (2014) presented cloneHD, a method that combines the use of copy number and somatic point mutations to infer the clonal population structure.

Subclones successfully deconvoluted can be used for phylogenetic reconstruction, which involves establishing evolutionary relationships between the clusters, ordering those clusters and rebuilding the tumour evolutionary trees (Niknafs et al. 2015) (Figure 1-3). Jiao et al. (2014) developed PhyloSub to infer relationships between the tumour clones using MCMC sampling. Similarly, Malikic et al. (2015) presented CITUP, a method for inferring phylogenies from multiple samples of the same patient. Both PhyloSub and CITUP reconstruct phylogenies based on somatic single nucleotide variants. Schwarz et al. (2014) and Zaccaria et al. (2017) also presented tools for deconvolution and phylogeny reconstruction based on copy number aberrations, MEDICC and CNT-MD, respectively. Deshwar et al. (2015) proposed PhyloWGS, a method very similar to cloneHD in that it combines both copy number and somatic point mutations, while also providing phylogeny reconstruction.

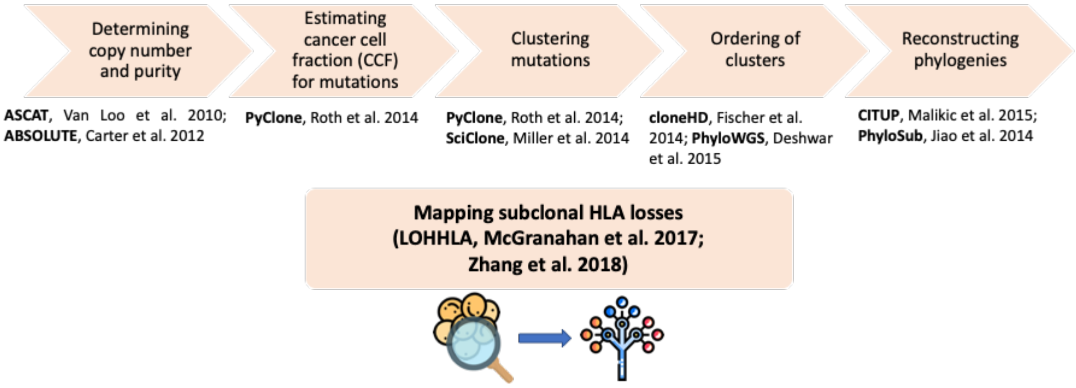

Determining copy number and purity
Estimating cancer cell fraction (CCF) for mutations
Clustering mutations
Ordering of clusters
Reconstructing phylogenies
ASCAT, Van Loo et al. 2010; **ABSOLUTE**, Carter et al. 2012
PyClone, Roth et al. 2014
PyClone, Roth et al. 2014; **SciClone**, Miller et al. 2014
cloneHD, Fischer et al. 2014; **PhyloWGS**, Deshwar et al. 2015
CITUP, Malikic et al. 2015; **PhyloSub**, Jiao et al. 2014
Mapping subclonal HLA losses (LOHHLA, McGranahan et al. 2017; Zhang et al. 2018)

Figure 1-3 Main steps involved in the workflow for studying ITH in the context of tumour evolutionary trees based on point mutations with example tools used for each analysis.

1.8 Challenges of studying intratumour heterogeneity

1.8.1 Tumour sampling bias

Sampling the tumour mass is the first step of analysing the intratumour heterogeneity underlying tumour evolution. The sampling process should involve careful consideration to ensure the inclusion of diverse subclonal populations and allow for a comprehensive assessment of the tumour's genetic and phenotypic heterogeneity. Notably, tumours consist of diverse subclonal populations of cells with distinct genetic make-up. In the past, bulk sequencing has been the common approach for many cancer studies (Nik-Zainal et al. 2012; Gerstung et al. 2020). A single sample of a tumour, obtained from bulk sequencing data, provides a limited perspective as it offers only a glimpse into the entire tumour’s evolutionary history. It is important to acknowledge that certain alterations can be strictly subclonal and may not be present in the specific part of the tumour that was initially sampled. As such, these alterations may be missed when relying on a single sample for analysis. Furthermore, an alteration may be present in every cell of the tumour sample, while not being present in the other parts of the tumour that were not sampled, thus being misclassified as clonal while being subclonal.

To mitigate these issues, recent studies have demonstrated how spatial sampling of multiple regions allows for the detection of a larger number of events, as well as more accurate clonality determination. Aforementioned NSCLC TRACERx 100 study provided compelling evidence that 76% of subclonal mutations would have appeared clonal through single-sample analysis (Jamal-Hanjani et al. 2017). Additionally, significantly more mutations were identified with multi-region WES compared to single-sample analysis. Conversely, longitudinal studies offer another approach to the tumour sampling bias by considering multiple samples from the same patient taken at different time points (Jiang et al. 2016; Chen et al. 2020; Boyle et al. 2020; 2021; Cindy Yang et al. 2021). Taking different tumour samples over time helps in identifying emerging subclonal populations and pinpointing clonal sweeps.

1.8.2 Variant calling and purity estimation

Another challenge in studying ITH is precise estimation of the mutational cancer cell fraction. Accurate determination of CCF requires correct differentiation between the inherited (germline) variants and variants acquired during tumour development (somatic) for accurate variant calling. The appropriate distinction between the two requires matched tumour and normal samples (Koboldt et al. 2012; Cibulskis et al. 2013). Conversely, absolute copy number estimation is hindered by the intricacies of purity estimation. Estimating the proportion of tumour cells within a sample and identifying the contamination from normal cells, requires matched normal samples and robust computational tools for accurate copy number analysis (Van Loo et al. 2010; Mermel et al. 2011; Carter et al. 2012).

1.8.3 Subclonal deconvolution

As outlined before, spatial multi-region and temporal multi-sample sequencing alleviate some of the sampling issues hindering inference of accurate tumour evolutionary history. However, studies performing bulk sequencing still require robust computational methods for dissecting the heterogeneity within the tumours to accurately depict the distinct subclonal populations within a heterogeneous tumour mass. Accurate subclonal deconvolution requires sequencing data of good quality and depth. With low coverage, alterations of low variant allele frequency (VAF) remain undetected, while low quality leads to misclassification of alterations into wrong subclonal populations. Furthermore, validation of the results remains challenging. Available approaches involve benchmarking on simulated data or independent validation using single-cell sequencing (sc-seq) or fluorescence in situ hybridization (FISH). Tools for subclonal deconvolution commonly rely on the assumption of the infinite sites model, in which each mutation can only occur at a unique site and cannot disappear (Kimura 1969). However, this assumption has some inherent limitations, as it does not account for the possible occurrence of revertant mutations, recurrent mutations and deletions of loci harbouring mutations (Roth et al. 2014; Miller et al. 2014). Furthermore, the infinite sites model assumes that tumour evolution occurs under neutral evolution, disregarding how certain alterations confer a growth advantage and act as a selective pressure for other events to occur.

1.8.4 Phylogenetic inference

Phylogenetic reconstruction is a key step in backtracking the tumour's evolutionary history. It involves determining the evolutionary relationships between different clones. Phylogenetic reconstruction often relies on two simple rules: the pigeonhole rule and the crossing rule. The pigeonhole rule states that two mutational clusters whose prevalence together exceeds 100% cannot be placed on independent branches of an evolutionary tree. The crossing rule forces the descendent clones to carry a lower cellular prevalence than their ancestors in every tumour region (Malikic et al. 2015). These two rules ensure higher accuracy of tree reconstruction.

Some of the main challenges of phylogenetic reconstruction include data quality, tree complexity, homoplasy and viability of multiple trees. Firstly, phylogenetic reconstruction assumes a perfectly accurate subclonal deconvolution where each cluster's CCF is correctly estimated and mutations are clustered correctly. However, due to the presence of noise in the data, achieving perfect accuracy in subclonal deconvolution is an exceedingly difficult task. Small errors can be amplified in phylogenetic reconstruction, especially when homoplasy, independently developed shared alterations, is present. Secondly, trees with a higher number of subclones are more complex and require more computational resources when applying both the pigeonhole and the crossing rules. Coupled with inaccurate subclonal deconvolution, the resources required for accurate reconstruction of all potential phylogenies can be very computationally intensive. Lastly, phylogenetic reconstruction based on tumour clones can often present multiple alternative and equally viable trees as potential solutions, while independent approaches for tree validation are still lacking.

1.8.5 Integration of multimodal data

Many cancer studies continue to rely on independent analysis of single types of events or, at most, the analysis of single-omics data. Nik-Zainal et al. (2012) identified the mutational processes in 21 breast cancers based on single-sample genomic data. Navin et al. (2011) used single-cell (sc) RNA sequencing to elucidate expression heterogeneity and identify subclonal populations of breast cancers. Landau et al. (2014) studied the transcriptomic heterogeneity of DNA methylation-based on genome-scale bisulfite sequencing in chronic lymphocytic leukaemias (CLL). All of the aforementioned studies focused on the analysis of subclonal events for studying intratumour heterogeneity. However, each utilised exclusively one type of data. Boyle et al. (2021) used longitudinal smouldering multiple myeloma patient data to examine the changes of subclonal CNA and mutations on phylogenetic trees between samples of the same tumour collected at different times. Recently, Martínez-Ruiz et al. (2023) attempted to integrate the genomic and transcriptomic data by studying the allele-specific expression (ASE) in NSCLC TRACERx 421 and classifying them based on whether they were copy number-dependent or copy number-independent. While this study performed phylogenetic reconstruction and attempted to map the ASE on a tumour branch, it should be noted that the results were still based on regional data, rather than truly deconvolving the regional data and identifying the subclonal events.

1.9 Tracking of subclonal events on tumour evolutionary trees

Integration of multi-omics data is crucial for providing a more comprehensive analysis of the subclonal population and evolutionary constraints underlying tumour progression. McGranahan et al. (2017) presented Loss of Heterozygosity in Human Leukocyte Antigen (LOHHLA), which integrated the copy number and the SNV data by tracking subclonal CNA of HLA on SNV-based trees. The focus of this research remained solely confined to genomics data. However, it served as a proof-of-concept study for mapping subclonal events on tumour evolutionary trees, thereby allowing for the tracking of evolutionary changes and the potential integration of multimodal data. Zhang et al. (2018) later utilised the LOHHLA tool to perform a similar mapping of subclonal HLA LOH events to study the tumour-infiltrating lymphocytes (TILs) in high-grade serous ovarian cancer (HGSC). Recently, Miura et al. (2022) presented PhyloSignare for tracking the somatic mutational processes on tumour evolutionary trees. This framework brought the study of mutational signatures to the subclonal level, potentially paving the way for integrating multi-omics data.

Existing tools for the integration of multi-omics data allow for a combination analysis of multimodal data. These tools combine the data at the sequencing level, relying on complex models to identify novel evolutionary patterns (Silverbush et al. 2019; Chatsirisupachai et al. 2021; Menyhárt and Győrffy 2021). However, interpreting the results of these complex models can be challenging without existing reference datasets for comparison and validation. Combining the multi-omics data at the sequencing level further requires additional single-omics analysis to provide the biological context needed for the appropriate interpretation of the results.

There is still an unmet need for tools for tracking subclonal events on tumour evolutionary trees that can integrate different types of multi-omics data using the output from state-of-the-art tools.

Studies have shown the need for methods utilising the results of subclonal deconvolution and phylogenetic reconstruction (Jamal-Hanjani et al. 2017; McGranahan et al. 2017; Zhang et al. 2018; Boyle et al. 2021; Frankell et al. 2023; Martínez-Ruiz et al. 2023). Studying the timing of subclonal events provides opportunities to identify novel evolutionary principles underlying tumour evolution. Accurate ordering of the different clusters enables the detection of new subclonal drivers driving tumour progression and therapy resistance. Furthermore, the ability to map mutational signatures and compare different types of events on tumour evolutionary trees would aid in identifying parallel evolution and mutual exclusivity of events on a subclonal level.

1.10 Aims and outline of the thesis

This introductory chapter has presented an overview of the literature on tumour evolution, clonal expansion, the data and existing methods for studying intratumour heterogeneity. Furthermore, it has highlighted the main challenges hindering the accurate reconstruction of cancer's evolutionary history. Additionally, this chapter has revealed that, while a few tools have been developed to study subclonal events along the tumour trees, these were designed to re-interpret the data and more generic approaches utilising the output of the state-of-the-art tools are still lacking. To address the existing gap in the research and to provide the research community with the ability to accurately track and integrate various types of subclonal events, this thesis aims to:

1) Develop a robust, automated, integrative method to accurately track various types of subclonal events on tumour evolutionary trees, and which provides a measure of goodness of fit of its results.
2) Provide extensive testing with benchmarking on simulated datasets, comparing the developed tool against an existing methodology of a similar function. This analysis will aim to demonstrate the superior performance of the developed tool, highlighting its benefits and acknowledging its limitations.

3) Validate the tool's functionality and potential capabilities by mapping and integrating different types of subclonal events on real datasets.

The structure of this thesis will follow the outlined aims. Chapter 2 will present the data and the methods used for analysis throughout this thesis. Chapter 3 will describe the development of the tool, encompassing a detailed discussion of the process employed for testing and evaluating different methods and strategies with the overarching aim of providing the tool with a measure of goodness of fit for the results. Chapter 4 will focus on the simulation of the different copy number events and benchmarking the developed methodology against another method of similar functionality. Chapter 5 will showcase the tool's integrative functionalities on real datasets, mainly the NSCLC TRACERx 100 dataset, by providing some insights into tumour's evolutionary dynamics. Finally, Chapter 6 will summarise the novel findings presented in this thesis, as well as discuss the limitations of this research and explore future opportunities that have emerged as a result of the developed method.

This work aims to provide the research community with a novel methodology for the integration of the multi-omics dataset to track the subclonal changes on the tumour evolutionary trees, deepening the understanding of evolutionary principles underlying intratumour heterogeneity.

Chapter 2 Data and Methods

2.1 Introduction

This chapter provides an overview of the data, bioinformatics, and experimental methods employed in this thesis. This project extensively utilised externally pre-processed data from previous studies (Jamal-Hanjani et al. 2017; McGranahan et al. 2017). This chapter describes data acquisition and processing, various methods and statistical tools utilised in the thesis, and outlines the experimental methods and their significance.

2.2 Data used in this thesis

2.2.1 NSCLC TRACERx 100 multi-region data

The collection and processing of the data, including subclonal deconvolution with mutation clustering, phylogeny reconstruction and estimation of segmented allele-specific copy numbers, were performed externally (Jamal-Hanjani et al. 2017). The resulting data comprised mutational CCF, cluster CCF, the tumour evolutionary tree and the copy number data for each patient.

Tumour samples from at least two regions, separated by 0.3 cm to 1 cm, were collected from 100 adult patients with non-small cell lung cancer (NSCLC) between stages 1A and 3A (with one patient classified as 3B) (Jamal-Hanjani et al. 2017). In this cohort, there were 38 women and 62 men, classified according to the tumour stage, type of cancer and smoking history. Most of the patients (n = 62) were sampled at earlier stages: stage 1A and stage 1B. 61 patients were diagnosed with lung adenocarcinoma, 31 with lung squamous cell carcinoma (LUSC) and another 7 patients with other subtypes of non-small cell lung cancer.

2.2.1.1 Whole exome sequencing

Patient samples were randomized, and whole exome sequencing was performed on an Illumina HiSeq machine. 327 tumour regions (323 primary and 4 lymph-node metastases) were sequenced in total, matched by 100 germline samples drawn from the whole blood of the patients (with a median sequencing depth of 426x). The data was aligned to the reference human genome (hg19).

Picard tools v1.107 (http://broadinstitute.github.io/picard) was utilised to clean, sort and merge files from the same patient region and to remove duplicate reads. Picard tools v1.107, GATK v2.8.1 (Van der Auwera and O'Connor 2020) and FastQC v0.10.1(https://bioinformatics.babraham.ac.uk/projects/fastqc/) were combined for quality control metrics. SAMtools mpileup v0.1.19 (H. Li et al. 2009) was used to locate non-reference positions in tumour and germline samples (Jamal-Hanjani et al. 2017). Somatic mutations were called using MuTect v1.1.4 (Cibulskis et al. 2013) and VarScan2 (Koboldt et al. 2009) in the multi-region sequencing data. Additional filtering was performed to improve the accuracy of variant calling and decrease the rate of false positives. Variants were considered true positive only when VAF was greater than 2% and the mutation was called by both VarScan2 and MuTect. For mutations that were called in one or more regions but not ubiquitously across all of them, VAF restrictions were reduced to VAF being equal to or greater than 1%, allowing for the identification of low-frequency variants that would otherwise have been missed. The annotation of the variants was done with Annovar (Wang, Li, and Hakonarson 2010) and COSMIC v75 (Forbes et al. 2015).

2.2.1.2 Subclonal deconstruction and phylogeny reconstruction

PyClone, a Bayesian clustering method (Roth et al. 2014), was used to cluster the mutations to determine their clonality and estimate their cluster CCF. The mutation CCF is the cancer cell fraction, or the frequency, of the mutations.

Under the infinite sites model assumptions, each cluster represents a monophyletic group, and the cluster CCF represents the proportion of cells harbouring that particular set of mutations, i.e. the proportion of cells in the lineage defined by these mutations. Cluster CCFs were estimated from the frequency or the CCF of the mutations corrected by purity and copy number.

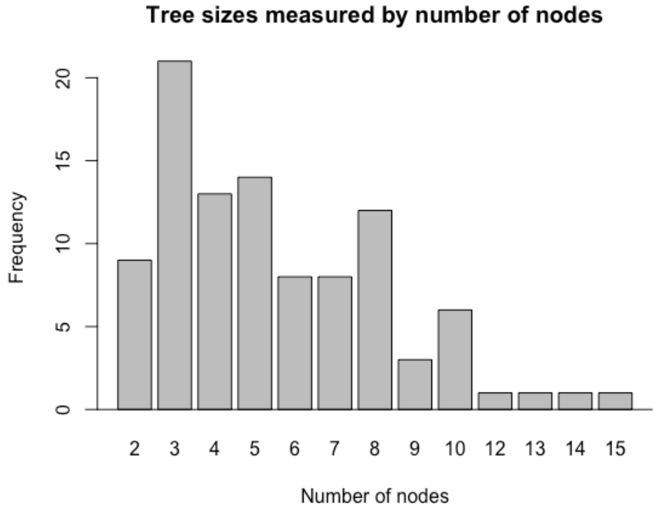
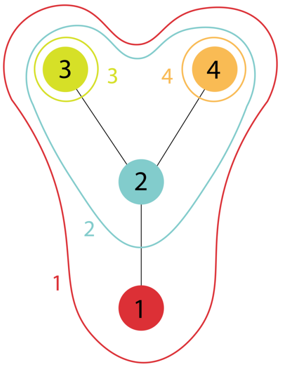


Figure 2-1 Frequency of different tree sizes measured by the number of nodes.

To provide higher accuracy of tree reconstruction, mutational clusters were first filtered based on the “pigeonhole rule” and “crossing rule”. The pigeonhole rule ensures that two mutational clusters whose prevalence together exceeds 100% cannot be considered independent and be placed on separate branches of an evolutionary tree, while the crossing rule states that a descendent clone must exhibit a lower cellular prevalence than its ancestor in every tumour region (Beerenwinkel et al. 2015). Only clusters with at least 5 mutations were included. CITUP was then employed for tree inference based on cluster CCFs, defining the relationships between different lineages (Malikic et al. 2015). In certain cases, CITUP identified several evolutionary trees as equally likely. For six patients, trees were constructed manually due to either number of clusters exceeding the maximum allowed in CITUP or erroneous copy number correction leading to trees and CCF values that were evolutionary nonsensical. Of note, tree sizes varied greatly between patients (Figure 2-1). Cluster CCFs were also used to estimate clone CCF – the proportion of cells with the same genotype, unique for that particular clone (Figure 2-2).

Cluster CCF

	Region 1	Region 2	Region 3	Region 4
1	100	100	100	100
2	80	90	95	85
3	40	20	35	40
4	30	50	30	15



Clone CCF

	Region 1	Region 2	Region 3	Region 4
1	20	10	5	15
2	10	20	30	30
3	40	20	35	40
4	30	50	30	15

Figure 2-2 The difference between cluster and clone CCF on an exemplary tree. The coloured area is cluster CCF and the coloured circles is the clone CCF. Cluster CCF includes the clone CCF with its descendants' CCF, while clone CCF is the proportion of cells with the genotype unique for that particular clone. Tables show the example Cluster CCF and Clone CCF for this particular tree.

2.2.1.3 Copy number analysis

The exome copy number data was processed with VarScan2 (Koboldt et al. 2012). The minimum coverage required was 8 reads. Homozygous and heterozygous single nucleotide polymorphisms (SNPs) were called in the germline using Platypus v0.8.1 (Rimmer et al. 2014) and then used to genotype the tumour regions. SNPs with coverage lower than 20x were filtered out. Log-ratio (LogR) and B-allele frequency (BAF) values were processed with ASCAT v2.3 (Van Loo et al. 2010) to provide the segmented allele-specific copy number data as well as the purity and ploidy estimates for all samples (Jamal-Hanjani et al. 2017). The ASCAT-inferred allele-specific copy number was corrected for purity.

2.2.2 HLA NSCLC TRACERx 100

Allele-specific HLA copy number data and HLA LOH events, classified based on their clonality in the NSCLC TRACERx 100 cohort, were obtained from McGranahan et al. (2017).

Firstly, the tumour and germline reads were extracted and mapped to the HLA locus using the SAMtools view. All unpaired reads were removed. The HLA typing was performed with POLYSOLVER (Shukla et al. 2015). SAMtools mpileup was used to calculate the coverage of the matched tumour and germline HLA alleles. Considering the polymorphic nature of HLA alleles, the HLA allele abundance was estimated by re-aligning candidate reads onto the HLA alleles inferred for each patient. The reads that mapped univocally on either allele in the tumour samples were counted and normalised against the coverage in the normal sample, determined with the R Biostrings package. HLA-specific coverage was determined at mismatch positions for pairs of homologous alleles ensuring accurate read counting. For reads that spanned more than one mismatch position, each read was counted only once to avoid duplication.

The LogR across each HLA gene was obtained by binning coverage across homologous alleles at intervals of 150 base pairs for both tumour and normal. Each bin was normalized by a multiplication factor M, corresponding to uniquely mapped reads for the germline, divided by uniquely mapped reads for the tumour. BAF was calculated as the coverage of one HLA allele divided by the sum of both HLA alleles. Finally, the HLA haplotype-specific copy number was determined utilising the LogR and BAF using the following equations:

$$CN_{Allele\ 1} = \frac{p - 1 + BAF \times 2^{LogR} \times (2(1-p) + p \times \psi}{p} \tag{1}$$

$$CN_{Allele\ 2} = \frac{p - 1 - 2(BAF - 1)^{LogR} \times (2(1-p) + p \times \psi}{p} \tag{2}$$

Where p was tumour purity and ψ was tumour ploidy input at the beginning. The BAF of the polymorphic site and the $LogR$ value found in the corresponding bin were used.

For each bin, the median Allele 1 and Allele 2 copy numbers were calculated as the median value across bins.

2.2.3 CRUK0640 from NSCLC TRACERx 421

The mutational data, copy number data, RNA reads, the phylogenetic tree, as well as the cluster CCF of patient CRUK0640, were obtained from the recently published NSCLC TRACERx 421 cohort (Frankell et al. 2023; Martínez-Ruiz et al. 2023).

For the RNA library preparation, FASTQ files were analysed using ARCHERDx analysis pipeline (v6.2.3) with default settings (https://archerdx.com/technology-platform/analysis/). The somatic mutation calling, copy number analysis and subclonal deconvolution methods were similar to the methodology previously described in the section for NSCLC TRACERx 100 with updated versions of the tools: SAMtools mpileup v1.10, VarScan2 v2.4.4, MuTect v1.1.7, GATK bundle v2.8, COSMIC v75, ASCAT v2.3, Platypus v0.8.1 and PyClone v0.13.1. The main changes involved using bam-readcount v0.8.0 for extraction of read information from the original alignment file for variant calling and using Sequenza v2.1.2 and ASCAT v2.3 for processing of the LogR data to provide somatic copy number aberration (SCNA) profiles (Van Loo et al. 2010; Favero et al. 2015). Lastly, the phylogenetic reconstruction was performed with a newly released method, CONIPHER (Grigoriadis et al. 2023), rather than CITUP. CONIPHER reconstructs phylogenies using the same filtering process as described in the section for NSCLC TRACERx 100 dataset. Clusters with their respective CCFs were used to reconstruct phylogenetic trees using the crossing and pigeonhole rules. CONIPHER removed spurious clusters, which were defined as clusters resulting from artefactual mutations or errors in SCNA calling. The errors of

SCNA were identified based on mutations co-localized in the genome. Then the tool attempted to reconstruct the phylogenetic tree following the two aforementioned rules while preserving the greatest number of mutations possible and removing clusters accordingly.

2.3 Methods

2.3.1 Simulations of CNA data with noise

Five distinct datasets, featuring different types of copy number alterations, were simulated to assess the performance of the different methods. These copy number alterations include loss of heterozygosity (CN 1>0), amplification (CN 2>3), duplication (CN 2>4), homozygous loss (CN 2>0) and heterozygous loss events (CN 2>1).

We used the patients' trees, mutation CCFs and cluster CCFs from NSCLC TRACERx 100 (Jamal-Hanjani et al. 2017) for the simulations. Patients lacking a tree, mutation CCF, and cluster CCF were excluded from the simulations, leaving 87 eligible patients. An event was simulated on every branch (including the root) of every patient, leading to 510 simulated events on 423 available branches and 87 roots across all 87 patient trees. Each simulation was performed by fitting a copy number value to the cluster CCF and patient tree across all regions. The equations for copy number fitting depended on the type of the event being simulated:

a) LOH events

$$CN_{sim} = 1 - \frac{CCF}{100} \tag{3}$$

b) Duplication events

$$CN_{sim} = \frac{CCF}{100} \times 4 + \left(1 - \frac{CCF}{100}\right) \times 2 \tag{4}$$

c) Amplification events

$$CN_{sim} = \frac{CCF}{100} \times 3 + \left(1 - \frac{CCF}{100}\right) \times 2 \tag{5}$$

c) Homozygous loss events

$$CN_{sim} = \left(1 - \frac{CCF}{100}\right) \times 2 \tag{6}$$

c) Heterozygous loss events

$$CN_{sim} = \frac{CCF}{100} + \left(1 - \frac{CCF}{100}\right) \times 2 \tag{7}$$

with constraints:

$$CN_{sim} \geq 0, CCF \in [0,100] \tag{8}$$

where $CN_{sim}$was the simulated copy number and CCF was the clone's cluster cancer cell fraction for a particular region.

Simulated noise was added to the fitted copy numbers. The noise around copy numbers was simulated using the following equation:

$$CN_{noise} \in \ unif\left(\max\left(0, CN_{sim} - \frac{noise}{100}\right), CN_{sim} + \frac{noise}{100}\right) \tag{9}$$

where CN_{noise} was the fitted copy number with simulated noise, $unif(x)$ was sampling from a uniform distribution, and $noise$ was the maximum noise range. The simulated control datasets included all noise ranges from 0 to 30 by increments of 5.

To assess the tools' performance on the simulated dataset, three different criteria were considered. Both tools were required to determine the correct clonality, the branch, and the copy number value for each simulated event to be considered correct. These criteria were nested within each other, as specified. The tool had to determine the correct clonality first to identify the correct branch. Likewise, for the correct copy number to be determined, both the correct branch and clonality had to be determined. One caveat is that MAPping SubClonal Events (MAPSCE) can provide more than one good solution. In such cases, the tool consolidates the solutions by calculating a consensus copy number state. For scoring purposes, a branch was considered correctly mapped if it was among the ones returned by MAPSCE. For the tool to accurately map the type of CNA of the simulated event, the difference between the inferred copy number by a tool and the simulated copy number had to be lower than 0.3 for every clone of the tree.

2.3.2 Gene annotation

Segmented copy number data was annotated using the Ensembl (release 104) gene annotation (Cunningham et al. 2022) accessed with biomaRt (Durinck et al. 2005; 2009). The driver gene datasets were obtained from the Molecular Signatures Database (Subramanian et al. 2005; Liberzon et al. 2011; 2015) and IntOGen (Martínez-Jiménez et al. 2020).

2.3.3 Gene sets testing

We tested gene lists for overrepresentation by comparing them against gene sets from Gene Ontology (GO) database (Young et al. 2010) and pathways from Kyoto Encyclopaedia of Genes and Genomes (KEGG) Database (Kanehisa et al. 2016). Both the GO gene set and KEGG pathway analyses were performed using the *goana* and *kegga* functions with default parameters from the limma R package (Law et al. 2014; Ritchie et al. 2015; Phipson et al. 2016).

2.3.4 Mutational Signature analysis

We estimated the relative contribution of mutational signatures on the subclonal level using the deconstructSigs R package (Rosenthal et al. 2016). The mutational signature analysis was performed only for the three patients of NSCLC TRACERx 100 in particular (CRUK0011, CRUK0068, CRUK0083). Five out of six tested lineages had at least 100 mutations, with the remaining one comprising 54 mutations. We analysed the samples only for the presence of mutational signature SBS3, which is associated with *BRCA1* and *BRCA2* mutations and whose proposed aetiology is the defective homologous recombination-based DNA damage repair. We used the default settings of deconstructSigs, which include the 27 reference mutational signatures from Alexandrov et al. (2013) and COSMIC v3 (Forbes et al. 2015).

2.3.5 dN/dS analysis

The dndscv R package (Forbes et al. 2015)(Martincorena et al. 2017) was used to quantify the selection for specific subclones of the NSCLC TRACERx 100 patients. The dN/dS values used in the analysis were the global maximum likelihood estimation (MLE) of the dN/dS (ω all), representing the variation of the mutation rate across genes (Martincorena et al. 2017).

2.3.6 Antigen presentation and processing genes

The antigen presentation and processing gene dataset was extracted from Gene Ontology (Ashburner et al. 2000; Aleksander et al. 2023). We mapped homozygous losses of the genes involved in the antigen presentation machinery (APM) to the trees of NSCLC TRACERx 100 patients. This allowed us to compare them against the HLA LOH events also mapped by MAPSCE. To further show the specific role of genes considered to be lost, the genes were tested for overrepresentation in the KEGG antigen presentation and processing pathway and visualized using R package, pathview (Luo and Brouwer 2013).

2.3.7 Statistical analysis

All statistical analyses were conducted in R (v4.0.0). Unless otherwise specified, all statistical tests were two-sided. For the association between the two groups, we used either Fisher’s exact test or Pearson’s chi-square test. Comparisons of distributions were performed using t-test.

Chapter 3 Mapping Subclonal Events development

3.1 Introduction

3.1.1 Intratumour heterogeneity at the core of tumour evolution

The introduction of driver mutations, which confer a selective advantage, gives rise to new subclones that form the core of the intratumour heterogeneity. ITH is an important cancer immune escape mechanism and a predictor of a patient's response to therapy (Hiley et al. 2014; Jiang et al. 2016; Li, Seehawer, and Polyak 2022)

Capturing the extent of ITH is not without challenges. Indeed, single-sample analysis provides an incomplete picture of the tumour's entire evolutionary history. Gerlinger et al. (2012) showed that 63 to 69% of all somatic mutations were not detectable across every tumour region in renal carcinomas. Jamal-Hanjani et al. (2017) stated that 86% of all tumour regions had region-specific subclones, and 65% of all subclones would have been identified as clonal in NSCLC, both of which emphasise the limitations of a single-sample analysis in accurately portraying the heterogeneity within a tumour mass.

Multi-region sequencing studies provide a more accurate picture of the subclonal populations driving ITH (Gerlinger et al. 2012; 2014; Jamal-Hanjani et al. 2017; Frankell et al. 2023). Timing of mutations and copy number alterations can reveal novel evolutionary mechanisms underlying tumour progression. In NSCLC, phylogenetic reconstruction allowed to identify potential parallel evolution of driver amplifications, including *RHOH*, *PHOX2B*, *BCL11A* and *CDK4* (Jamal-Hanjani et al. 2017). Determining the clonality of certain events allows for the classification of certain drivers based on their role in either tumour initiation, progression or maintenance. Alterations in *EGFR*, *MET*, *BRAF*, and *TERT* for adenocarcinomas, in *NOTCH1*, *FGFR1* for squamous-cell carcinomas, and in *TP53* for both lung cancer types, always appeared to be early clonal events occurring before genome duplication, suggesting a role in tumour initiation. Mutations of *KMT2C* and *COL5A2* in adenocarcinomas, and *PIK3CA* in squamous-cell carcinomas were predominantly clonal, but occurred after genome duplication, suggesting their role in tumour progression or maintenance (Jamal-Hanjani et al. 2017).

State-of-the-art methodologies focus on subclonal deconvolution and phylogenetic reconstruction (Roth et al. 2014; Miller et al. 2014; Malikic et al. 2015; Deshwar et al. 2015). However, most cancer studies have primarily focused on the analysis of a single type of alteration, or at most using single-omics data on tumour evolution (Nik-Zainal, Van Loo, et al. 2012; Gerlinger et al. 2012; de Bruin et al. 2014; Jamal-Hanjani et al. 2017), missing the bigger picture of how multiple alterations drive tumour evolution. Studying single-omics datasets helps in understanding the mechanisms driving clonal expansion. Single-omics data has allowed for the study of mutational signatures along the tree (Miura et al. 2022), classification of the clonality, the timing and thus the role of the drivers in tumour evolution (Boyle et al. 2021), identification of evolutionary dynamics such as parallel evolution (Jamal-Hanjani et al. 2017), and identification of novel mechanisms of immune escape (McGranahan et al. 2017).

However, there has been little to no effort to integrate the multi-omics data at the subclonal level, including copy number, gene expression and methylation changes on the tumour evolutionary trees.

McGranahan et al. (2017) presented LOHHLA, a tool for the estimation of allele-specific HLA loss from sequencing data. This study revealed the effects of HLA LOH on early-stage NSCLC. Loss of heterozygosity of HLA locus occurred in 40% of lung cancer patients. Furthermore, subclonal cases of HLA LOH could be related to a higher non-synonymous mutation rate and neoantigen burden (McGranahan et al. 2017). LOHHLA used quadratic programming, an optimisation method, in a branch test to map the subclonal copy number changes on a particular branch of the tumour evolutionary tree (McGranahan et al. 2017). LOHHLA was the first tool that allowed the integration of mutation and copy number data through mapping subclonal events on SNV-derived tumour trees.

Since then, Zhang et al. (2018) used a similar approach to study tumour evolution in ovarian cancer, where they showed that subclonal HLA LOH was linked to higher CD8+ TIL levels. They developed their own methodology to map subclonal HLA LOH events on SNV-based tumour evolutionary trees. This approach involved the use of Bayesian Inference and MCMC (Zhang et al. 2018). These methods for mapping subclonal HLA LOH could be extended to include other types of copy number alterations and to integrate multi-omics data in general.

3.1.2 Methodology in LOHHLA

LOHHLA allows for the detection of HLA allele losses, as well as direct mapping of the subclonal cases on the corresponding SNV-based patient evolutionary tree (McGranahan et al. 2017). LOHHLA improves the CN detection of the highly variable HLA loci by identifying the HLA alleles for a particular patient and remapping the sequencing reads on these alleles to quantify their CN states. LOHHLA determines the clonality of the event based on the inferred CN states and in cases of a subclonal LOH event, it uses quadratic programming (QP) to map this event on the patient tumour trees.

3.1.2.1 Clonality determination

For clonality determination, LOHHLA uses the observed copy number values, where these are inferred CN states of the HLA alleles. In cases where the observed copy number values across all samples are lower than 0.5, LOHHLA considered the allele to be clonally lost. If the observed copy number values across all the regions are all higher than 0.5, LOHHLA would determine that there was no LOH at all. Lastly, for cases where observed copy numbers were varied, with some lower and some higher than 0.5, LOHHLA would determine the allele to be subclonally lost. Only after this filtering step, LOHHLA performs a branch test to map the LOH subclonal event.

3.1.2.2 Branch test

LOHHLA transforms the cluster CCF, the proportion of cancer cells in a particular lineage, onto the clone CCF, the proportion of cancer cells with the same genotype unique for that particular clone. Assuming that there has been one and only one subclonal copy number event, LOHHLA performs a branch test, splitting the tree at any particular branch to estimate the copy number state before (CN_{before}) and copy number state after (CN_{after}) the branch. The proportion of mutations in both parts of the tree can be estimated by summing the clone CCFs in the lineage after the branch or in the rest of the tree. This problem can be written as a system of *n* equations with two unknowns, where *n* is the number of regions and each equation represents the relationship between the observed, empirical copy number values for that particular region and the linear combination of the clone CCFs multiplied by the CN_{after} or CN_{before} depending on whether the clone appears within the lineage defined by the branch or not. The expectation is that a subclonal, allele-specific copy number loss would be detected as $CN_{before} = 1$ and $CN_{after} = 0$ when considering the branch where the loss occurred.

The branch test aims to identify which branch best explains the observed data as a LOH event.

3.1.2.3 Quadratic Programming

The branch test in LOHHLA utilises quadratic programming to solve the aforementioned system of equations. Quadratic programming in LOHHLA's branch test was used to find the best matching copy number states before and after each branch of the tree under certain constraints, in this case, both solutions must be non-negative. In LOHHLA's branch test, the copy number state before the event was constrained to be larger than 0.5. The solution to the system of equations can be presented as a quadratic programming equation (Equation 10):

$$\min\left(-d^T b + \frac{1}{2} b^T D b\right) \tag{10}$$

under constraints:

$$A^T b \geq b_0$$

where *d* and *b* are *n*-vectors, *D* is a *2n* symmetric positive definite matrix, *A* is an *n x m* matrix and *b₀* is an *m*-vector (Goldfarb and Idnani 1982; 1983).

QP can naturally limit the solutions to non-negatives and is therefore ideally suited for resolving this kind of problem since negative CN values are non-sensical. Since the branch test in LOHHLA was designed specifically to map subclonal LOH events of the HLA allele, it also constrained the inferred copy number value to be at least 0.5 at the root of the tree.

In this thesis, LOHHLA's mapping approach (LOHHLA/QP) refers to these last two parts of the method, namely the clonality determination step and the branch test, involving the mapping of subclonal events on the tumour evolutionary trees.

The branch with the lowest sum of squares of residuals (RSS) is the best-fitting model (Equation 11). However, the main problem of this methodology is the lack of uncertainty quantification.

$$\mathrm{RSS} = \sum\left(\mathrm{CN}_{\mathrm{observed}} - \mathrm{CN}_{\mathrm{predicted}}\right)^2 \tag{11}$$

Other methods that could provide uncertainty measurement for the results of the branch test were also explored in this thesis.

3.1.3 Limitations of LOHHLA

One of the major limitations of LOHHLA's mapping approach was the lack of uncertainty quantification and the lack of measurement of goodness of fit for the results beyond the RSS. More precisely LOHHLA/QP does not have any indication on whether the best result was good enough or whether the second or third results were qualitatively just as good as the first one. In addition, the threshold approach to determining clonality did not allow for accurate measurement of the clonality determination itself. Indeed, LOHHLA's mapping approach did not contrast the results with the possibility of not having any subclonal event (i.e. null hypothesis). Lastly, the parameters for QP in LOHHLA's mapping approach were specifically selected for subclonal LOH events and thus used rigid constraints that limit the approach to only copy number losses. A more flexible approach would allow for the mapping of copy number gains or other data types like expression values, for instance.

3.1.4 Mapping other subclonal copy number events and integration of multi-omics data

This study provided the opportunity to extend this approach to map different types of subclonal copy number, gene expression and methylation changes on the tumour evolutionary trees, beyond HLA allele losses only. Such an approach would have to be able to provide an estimate of the uncertainty in the results to help with their interpretation. This project focuses on the development of a methodology specifically for mapping subclonal events, initially for both copy number gains and losses, but that can be extended to map other data types.

The majority of this PhD project were committed to exploring different potential approaches to mapping subclonal events on the tumour evolutionary trees, and the development of the methodology, including simulating noise around the CCF for assessing the soundness of these approaches. This chapter describes the work and reasoning behind the features included in the release of the tool on GitHub (v1.0.0), and the utilisation and validation of the different functionalities.

3.2 Results

This section provides insight into how the tool was developed and why certain features were included in the release of the tool on GitHub (v1.0.0).

3.2.1 Non-negative least squares

Non-negative least squares (nnls) is a form of least squares where coefficients are constrained to be non-negative. Least squares is an optimisation approach where the solution minimizes the sum of the squares of the residuals. Residuals are the differences between the observed values and the values fitted by the model (Figure 3-1). Note that squaring the differences leads to a bias towards larger outliers.

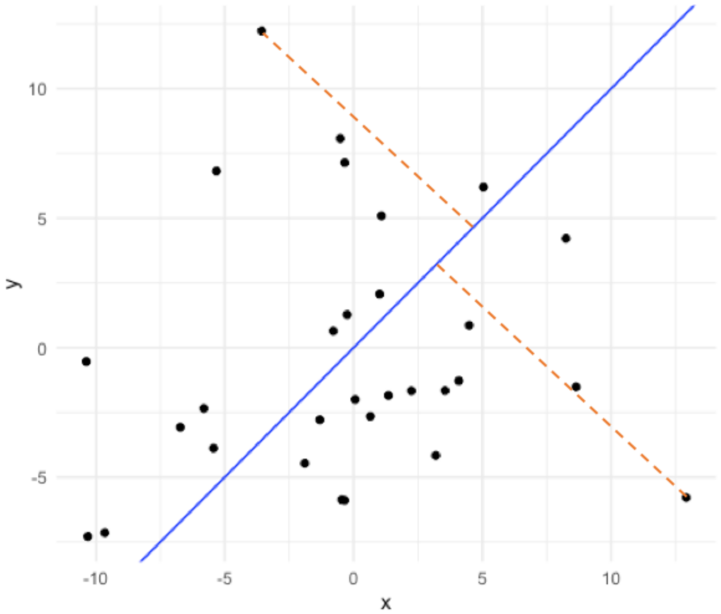


Figure 3-1 Least squares approach. Black points are data points, the blue line shows the fitted model, while the red dashed line indicates the

difference between two observed values and two values fitted by the model.

We explored non-negative least squares optimisation for the branch test. The approach was implemented using the *nnls* R Package (Stokkum 2012). In all cases, the results of running the branch test with non-negative least squares and quadratic programming were identical. Both tools identified HLA LOH events on the same branches with the same error for patient CRUK0098 (Figure 3-2).

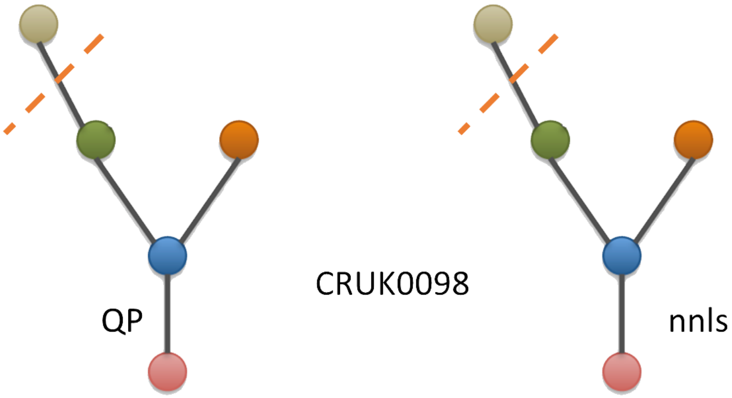


Figure 3-2 Comparison of the quadratic programming (QP, on the left) and non-negative least squares (nnls, on the right) used for branch testing on patient CRUK0098. The dashed line indicates the branch where HLA LOH was detected.

Non-negative least squares is an equivalent of the quadratic programming approach currently used in the branch test. This is because non-negative least squares minimisation:

$$\arg min_x \left||Ax - y|\right|^2 \tag{12}$$

under constraints:

$$x \geq 0$$

where *A* is an *n x m* matrix, *x* is an *n*-vector and *y* is an *m*-vector (Bro and Jong 1997), can be presented as a form of quadratic programming minimisation:

$$\arg\min(\frac{1}{2}x^TQx + c^Tx) \tag{13}$$

under constraints:

$$x \geq 0$$

where:

$$Q = A^TA \; and \; c = -A^Tx \tag{14}$$

Thus, non-negative least squares was equivalent to QP and therefore provided no advantage.

3.2.2 Deriving CCF from observed copy number values

We also explored deriving cancer cell fractions from observed copy number values for the purpose of mapping events on the tumour evolutionary trees. The CCF values can be derived from copy number for each region and mapped on the cancer evolution tree by re-clustering the somatic mutations with copy number CCF into cluster CCFs. This could provide explicit information regarding copy number alterations in the clustering step and present an additional validation of quadratic programming and mutation clustering. The conversion of the observed copy number to CCF was performed using two different strategies, in an attempt to integrate the observed CN values for the HLA alleles into the initial clustering of the SNV CCFs.

The first approach focused on HLA LOH cases detected by LOHHLA where the observed copy number was 1 in one region and lower than 1 in another. In these cases, an assumption was made that a loss happened between the ancestral state ($CN_{before} = 1$) and the descendant state ($CN_{after} = 0$). Thus, the CCF derived from CN would be the CCF of the loss denoted by the percentage of copy number lost between the two regions. This can be represented by the equation:

$$CCF_{CN} = 1 - CN_R \tag{15}$$

where:

$$CN_{R1} = 1$$

$$0 < \; CN_{R2} < 1$$

thus:

$$CCF_{R1} = 0$$

$$CCF_{R2} = 1 - CN_{R2}$$

where CCF_{CN} denotes the CCF derived from CN_{obs}, CN_R is the regional copy number, CN_{R1} is copy number for one region and CN_{R2} is copy number for the other region. CCF_{CN} ranges from 0% when the CN_{R2} is 1 to 100% when the CN_{R2} is 0.

This approach provided straightforward solutions in these simple cases. Visually, one can display the CCF_{CN} onto the scatterplot with all mutation CCFs between two regions such that the CCF_{CN} will always cluster with the mutation CCFs defining the clone where HLA LOH event was detected (Figure S3-1, Figure 3-3). However, this method is only applicable to cases where the CN_{before} and CN_{after} are known, hence it was not possible to generalise this procedure to all types of copy number events without having an orthogonal method to infer both the ancestral and derived states.

The second approach to deriving CCF from CN utilised the copy numbers estimated by quadratic programming. For consistent conversion of CN to CCF values, the following equation was used:

$$CCF_{CN} = \frac{|CN_{obs} - CN_{QP.before}|}{|CN_{QP.after} - CN_{QP.before}|} \tag{16}$$

under constraints:

$$CN_{QP.after} \neq CN_{QP.before}$$

where CCF_{CN} is the cancer cell fraction for the copy number event, CN_{obs} is the observed copy number value in a given region, and $CN_{QP.before}$ and $CN_{QP.after}$ are the quadratic programming-inferred copy number states before and after the branch, respectively.

The resulting CCF_{CN} was then mapped onto a 2-dimensional scatter plot of all mutation CCFs between two regions. In most cases, the CCF_{CN} was located close to the cluster of mutations that denoted the branch where the LOH event occurred.

Utilising copy number estimated by quadratic programming for conversion of CN_{obs} into CCF_{CN} (Equation 16) made it possible to extrapolate the results for all patients regardless of the ancestral state. However, this approach required the use of QP to infer $CN_{QP.before}$ and $CN_{QP.after}$ and therefore was not sufficient on its own. In one example case, the CCF_{CN} for patient CRUK0039 pointed exactly towards the cluster, where HLA LOH was detected from branch testing with quadratic programming (Figure 3-3, panels A-B). This result occurred for most cases except in a few exceptions. For instance, in patient CRUK0005 the CCF_{CN} was placed visibly away from the target cluster in R3 (Figure 3-3, panels C-D). This could have happened in instances where the noise levels in the data were too high. The example case of patient CRUK0005 demonstrate there was no cluster with a high CCF in R1, R2 but low in R3. Cluster 3 seems to be the best approximation for the subclonal HLA LOH event, considering the placement of the derived CCF_{CN} close to the cluster 3,

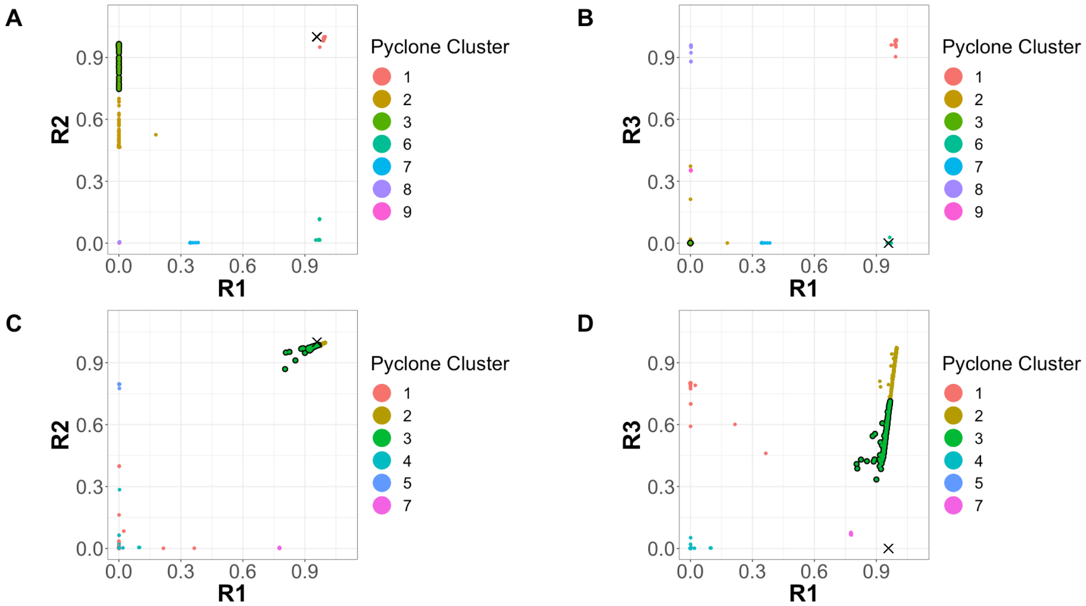


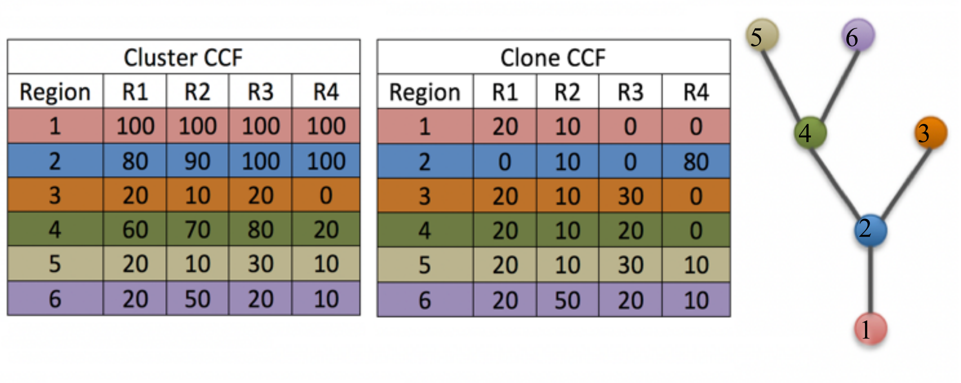
Figure 3-3 Mutation CCFs, HLA LOH cluster and CCF_{CN}. Region 1 (R1) vs region 2 (R2) (A) and R1 vs region 3 (R3) (B) for patient CRUK0039, respectively. R1 vs R2 (C) and R1 vs R3 (D) for patient CRUK0005, respectively. Mutations were coloured by their assigned cluster during subclonal deconvolution using PyClone, labelled as Pyclone Cluster in the legend. The cluster where HLA LOH was detected is highlighted by a thicker outline of the cluster. The black cross indicates the derived CCF from copy number (CCF_{CN}). The placement of the black cross close to a particular cluster denotes denotes a potential mapping of the *CCF*$_{CN}$ on that cluster.

3.2.3 Statistical support for the branch test

In addition to assessing alternatives to quadratic programming, several methods to provide additional statistical support to the results were explored.

3.2.3.1 Simulating noise in an artificial dataset

An artificial dataset was simulated for testing the different approaches for uncertainty measurement. As a simple approach to simulate the noise in real data, a random number was picked from a uniform distribution for different noise ranges, from ±1% to ±10%, in increments of 1%. The noise was then added to each of the cluster CCFs. 10,000 simulations were performed for each noise level. An LOH event was simulated in four different branches from the tree depicted in Figure 3-4 (branches 3, 4, 5 and 6) and the $CN_{observed}$ values for each region were derived.



Cluster CCF				
Region	R1	R2	R3	R4
1	100	100	100	100
2	80	90	100	100
3	20	10	20	0
4	60	70	80	20
5	20	10	30	10
6	20	50	20	10

Clone CCF				
Region	R1	R2	R3	R4
1	20	10	0	0
2	0	10	0	80
3	20	10	30	0
4	20	10	20	0
5	20	10	30	10
6	20	50	20	10

Figure 3-4 Cluster and clone CCF and tumour tree made for the artificial dataset, coloured by clone.

As expected, the percentage of correct solutions degraded with the increase in noise (Figure 3-5). For each tested branch, adding noise of up to ±10% led to the lowest percentage of correct answers, while with noise up to ±2 all results were still correct.

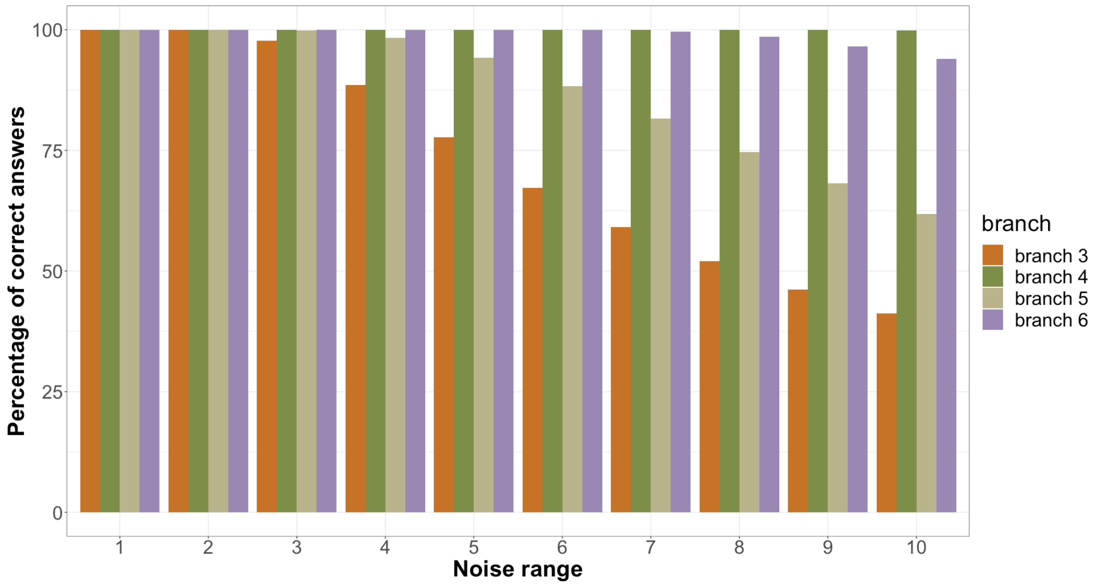


Figure 3-5 Percentage of correct results for different maximum noise ranges. Each bar represents 10,000 simulations. Different colours indicate the branch where the LOH event was simulated.

Interestingly, the noise increase did not have the same consequence in all 4 branches. Results for branch 3 were the most affected by noise, with the percentage of correct results declining to less than 50% for the highest level of tested noise. Conversely, simulated events on branches 6 and 4 seemed more robust to noise, with the percentages of correct answers always over 90% for branch 6 and staying at 100% for branch 4. The explanation for this lies in the size of the lineages. While clusters 3 and 5 never represented more than 30% of any region, clusters 4 and 6 reached at least 50% in one of the regions (Figure 3-4), resulting in a higher signal-to-noise ratio, even in the presence of higher noise levels.

3.2.3.2 Maximum Likelihood Estimation and Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) (Sunnåker et al. 2013) is a computational approach for estimating posterior distributions of model parameters using Bayesian statistics. In this case, ABC can be used to compare the posterior probability for each branch.

In an attempt to employ ABC, we utilised an artificial dataset with simulated noise around CCF (Figure 3-4). The best-fitting CN states before and after each branch of a tree using QP were calculated to provide an RSS for each possible solution. The cluster CCFs were re-built by incorporating simulated noise ranging from 5 to 10%. The noise was sampled from a uniform distribution. The CCF values were truncated to fit in the range between 0 – 100% CCF. Approximate Bayesian Computation was then used to estimate the posterior probabilities and the goodness of fit of each model (branch of the tree). This was done on simulated data for 10,000 simulations, which resulted in a pseudo-ABC approach that resembled maximum likelihood estimation. Implementation of pseudo-ABC yielded posterior distributions for different models, which corresponded with the percentage of correct answers obtained by running the branch test with quadratic programming. However, this approach was very computationally intensive and other options were considered.

3.2.3.3 Exploring bootstrapping

Bootstrapping is a test which relies on sampling with a replacement that permits estimating the stability of a particular solution. Therefore, it provides a measure of the uncertainty in the result. In this case, the aim was to recalculate the cluster CCFs by sampling from the mutations in each cluster. After each sampling, QP was used to infer the best matching branch, CN_{before} and CN_{after}, and obtain the corresponding RSS. Because of the resampling, the bootstrapped RSS values were an average and therefore more robust than the RSS inferred from a single branch test. Unlike pseudo-ABC or ABC, bootstrapping was more efficient and required fewer computational resources.

Bootstrapped RSS values could be used to calculate a Bayesian Information Criterion (BIC) and Bayes Factors (BF) for model selection.

3.2.3.4 Filtering of results

In the initial stages of the methodology development, the QP-based branch test identified multiple potential results for every allele. These results underwent a filtering process to remove spurious subclonal events. Using quadratic programming-based branch test with bootstrapping, 2056 putative CN changes were found for the NSCLC TRACERx 100 HLA dataset (McGranahan et al. 2017; Jamal-Hanjani et al. 2017) (Figure 3-6). These CN changes included cases, where the branch test identified a couple of potential events per HLA allele.

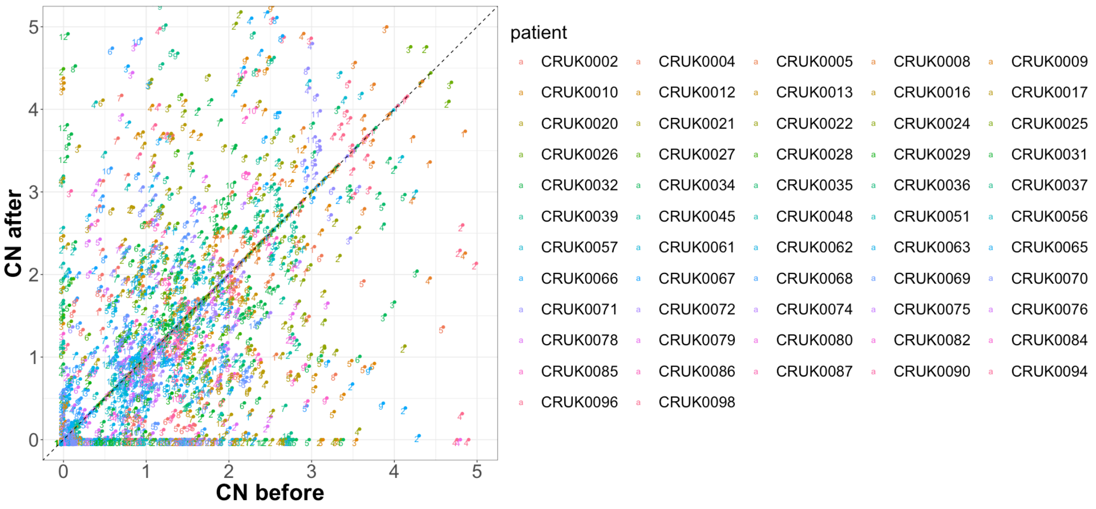


Figure 3-6 All 2056 putative HLA copy number events for all alleles in every patient. X-axis shows inferred copy number before the event, and Y-axis shows inferred copy number after. Each point denotes a branch, coloured by the patient. The diagonal (dashed line) corresponds to $\mathrm{CN}_{\mathrm{before}} = \mathrm{CN}_{\mathrm{after}}$.

For each branch test, an RSS was obtained, which was then converted to BIC for model selection. By bootstrapping the cluster CCFs 100 times, we generated 100 BIC values to be averaged for every putative CN change. The first filter involved selecting branches with a significant difference between the BIC means of the branches. This statistical test was performed using a one-way analysis of variance (ANOVA) to check for significant differences between the BIC means of all the branches for each HLA allele. Additionally, only results with at least strong evidence for the most fitting BIC, compared to the next best one, were considered (within a difference of 6 Bayes Factors after conversion from BIC, as described in the later section of this chapter). Lastly, only branches, for which at least 95% of their bootstrapped BIC values were better than the BIC of the null hypothesis were included, resulting in 720 detected subclonal HLA allele CN changes over 57 patients (Figure 3-7).

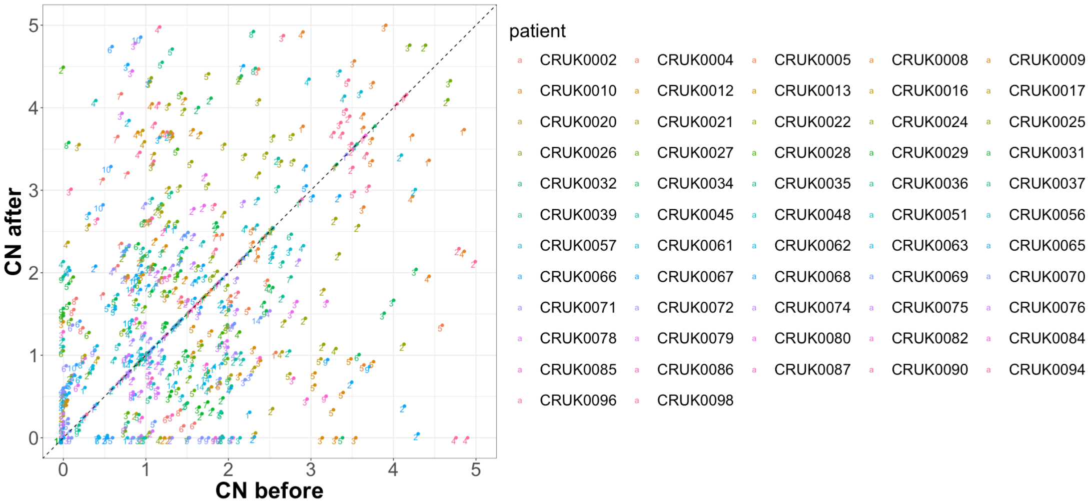


Figure 3-7 720 HLA CN changes (CN_{before} vs CN_{after}) after applying the statistical filters. X-axis shows inferred copy number before the event, and Y-axis shows inferred copy number after. Each point denotes a branch, coloured by the patient. The diagonal (dashed line) corresponds to $CN_{before} = CN_{after}$.

Finally, for detection of HLA LOH events, branches were filtered to only include solutions where CN_{before} is larger than CN_{after} by at least 0.5 for smaller CN_{before} (at most 1), and by at least 1 for larger CN_{before} (at least 2). This yielded only 68 cases of subclonal HLA LOH events in 33 patients (Figure 3-8).

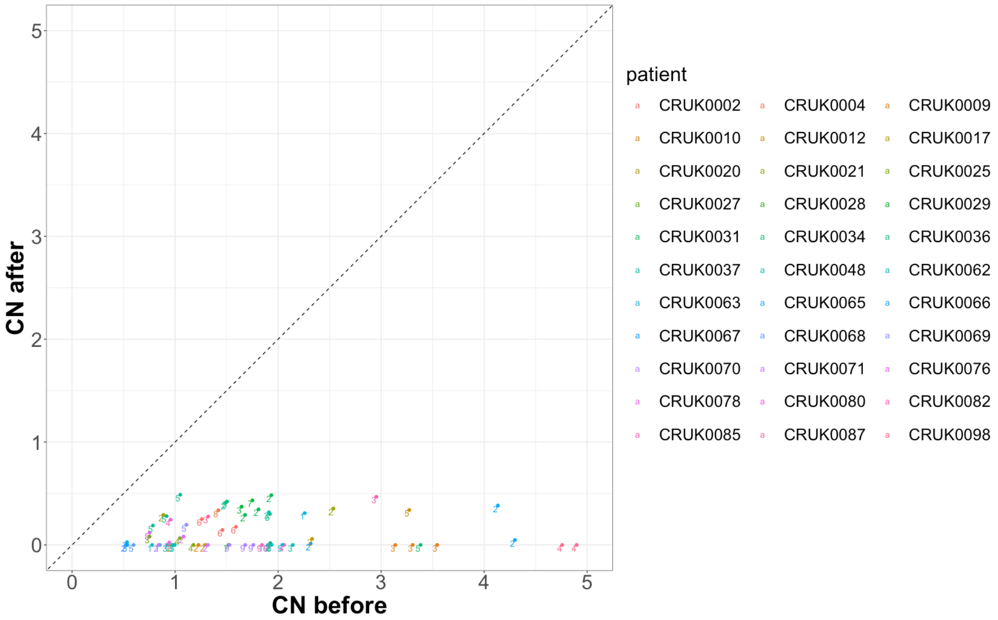
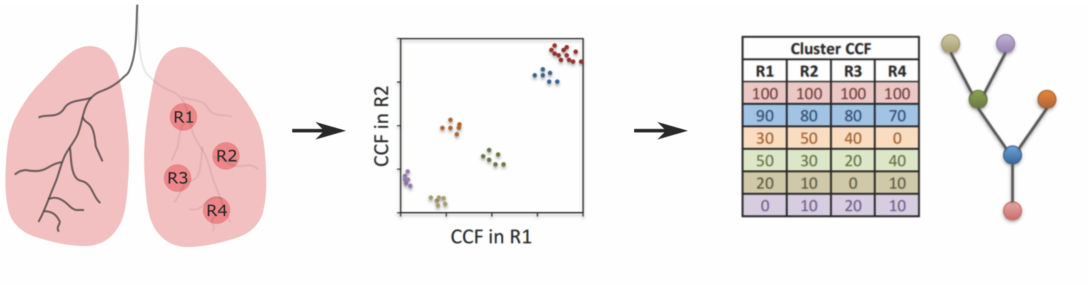


Figure 3-8 CN cases of subclonal HLA losses in 33 patients. X-axis shows inferred copy number before the event, and Y-axis shows inferred copy number after. Each point denotes a branch, coloured by the patient. The diagonal (dashed line) corresponds to CN$_{\text{before}}$ = CN$_{\text{after}}$.

3.3 MAPping SubClonal Events method

3.3.1 Overview of the tool

MAPping SubClonal Events (MAPSCE) is a computational approach designed to map subclonal events on tumour evolutionary trees. It simultaneously infers the clonality of an event and, in the case of a subclonal event, maps it on the tree while providing robust measurements of goodness of fit. It leverages multi-region sequencing data. MAPSCE clusters the mutations based on their cancer cell fraction to calculate the cluster CCF. Using the cluster CCF, the mutational CCF, the tumour evolutionary tree and the observed data (Figure 3-9), MAPSCE performs a branch test using quadratic programming on every branch of the tree in turn, by solving the system of equations (Figure 3-10). The method assumes that, for a particular gene or locus, at most one subclonal event can take place. This assumption allows for more straightforward analysis with higher computational efficiency. However, this assumption reduces the complexity of the biological data, which sometimes may involve repeated alterations affecting a particular gene or locus. The tool resamples the cluster CCF and runs the branch test 100 times, from which it derives an average bootstrapped RSS value for each branch, and the null hypothesis (the root), to assess how well a clonal event matches the observed CN values. The bootstrapped RSS values can then be converted into Bayesian Information Criterion for comparison of the distributions of the BIC for different branches and the null hypothesis (the trunk), and the BIC values can be, in turn, converted into Bayes Factors. MAPSCE uses the Bayes Factors for model selection (where each branch is a model) by assessing the strength of the evidence for each branch against the top-scoring one. In cases where more than one good solution is found, MAPSCE provides a consensus average copy number state for the clones where the solutions agree with each other. Lastly, where only two regions are available, the tool can automatically run without bootstrapping to avoid adding more noise to the data. Each feature of the tool is described in detail in the following sections.

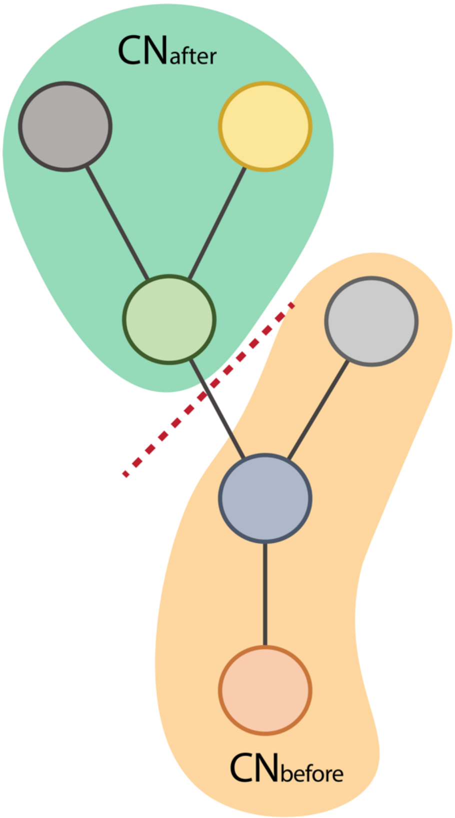


Cluster CCF			
R1	R2	R3	R4
100	100	100	100
90	80	80	70
30	50	40	0
50	30	20	40
20	10	0	10
0	10	20	10

Figure 3-9 The tool relies on the use of multiregion sequencing data, i.e. mutational and cluster CCF, as well as the tumour evolutionary tree.

3.3.2 Quadratic Programming

As previously described, the tool utilises quadratic programming for its branch test to find the optimal solution for every branch of the tree and finally choose the most likely branch or branches where the event might have happened (Figure 3-10). Unlike LOHHLA/QP, where the copy number states are constrained to CN_{before} being at least 0.5, and the CN_{after} being non-negative, the default settings in MAPSCE constrain CN_{before} and CN_{after} to be non-negative. However, the tool has been designed to allow the user to set custom constraints on CN_{before} and CN_{after} for cases where the type of CN event or any other event has been pre-determined in the dataset.



Copy number			
R1	R2	R3	R4
0.4	0.3	0.5	0.8

$$R1: 0.4 = 50CN_{before} + 50CN_{after}$$
$$R2: 0.3 = 30CN_{before} + 70CN_{after}$$
$$R3: 0.5 = 40CN_{before} + 60CN_{after}$$
$$R4: 0.8 = 80CN_{before} + 20CN_{after}$$
$$\rightarrow CN_{before} = 0.73 \;\&\; CN_{after} = 0$$

Figure 3-10 MAPSCE's branch test uses the observed data, i.e. patient tree, the absolute copy numbers, and the cluster CCFs, and can be represented as a system of equations with two unknowns. The value CN_{before}represents the CN state before the tree and CN_{after} represents the CN state after the tree. The coloured circles represent the clones of the tree (the colours of the circles are arbitrary), while the coloured areas denote the CN states before (orange) and after (green) the chosen branch.

Quadratic programming is used to solve the system of equations, where each line of equations represents a different region of the dataset. The left side of the equation represents the observed data, copy number in the example provided, while the right side of the equation estimates the proportion of mutations of two parts of the tree, the cluster CCF of the lineage before (ancestral state) and after (derived state) the branch where the event happens. This is a system of n equations (where n is the number of regions) with two unknowns, the copy number value before the branch (CN_{before}) and the copy number state after the branch (CN_{after}):

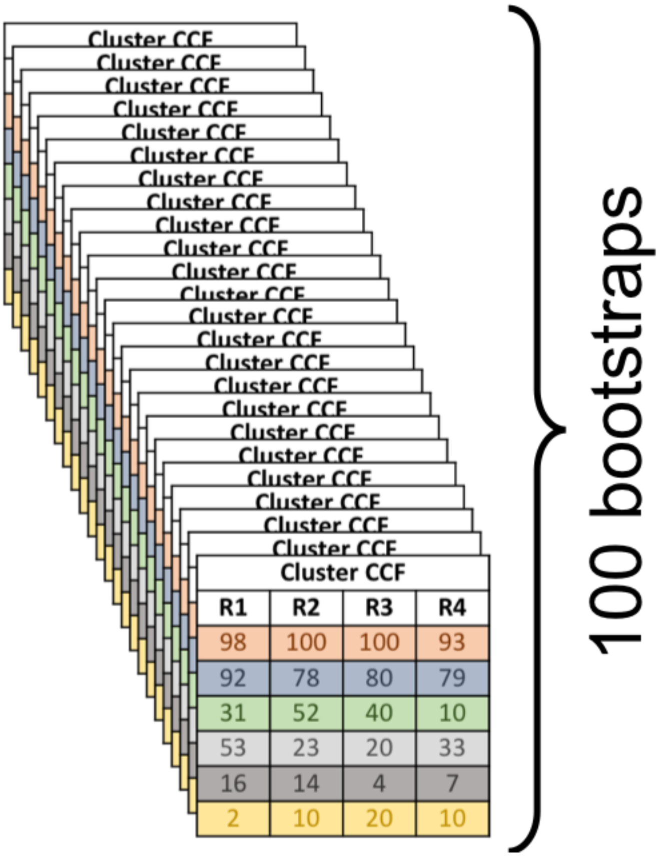
$$\begin{gathered} R_1: CN_{obs_1}{=}CCF_{before_1} \times CN_{before} + CCF_{after_1} \times CN_{after} \\ R_2: CN_{obs_2}{=}CCF_{before_2} \times CN_{before} + CCF_{after_2} \times CN_{after} \\ \vdots \\ R_n: CN_{obs_n}{=}CCF_{before_n} \times CN_{before} + CCF_{after_n} \times CN_{after} \qquad (17) \end{gathered}$$

CN_{obs} is the observed copy number, CCF_{after} is the cluster CCF of the lineage after the branch and CCF_{before} refers to the rest of the tree, in this case $1 - CCF_{after}$, which can be calculated by subtracting CCF_{before} from 100 or subtracting 100 from CCF_{before} in cases where the CCFs erroneously add up to more than 100%.

For every QP solution, an RSS is calculated (Equation 11) where a lower RSS represents a better fit. However, MAPSCE converts the RSS into BIC first and then into Bayes Factors for scoring the branches.

3.3.3 Bootstrapping

As previously described, bootstrapping is a resampling technique where sampling is performed with replacement, and it can be used to provide a measure of accuracy to a particular estimate. In MAPSCE, bootstrapping is performed by resampling mutations from each cluster independently. In each case, the cluster CCF is recalculated as the average CCF of the sampled mutations. This allows MAPSCE to explore the uncertainty in the cluster CCFs and how that affects the mapping of the subclonal events. By default, 100 bootstraps are performed per branch and an RSS for every bootstrap is available, summarised as a mean RSS for each branch (Figure 3-11).



R1	R2	R3	R4
98	100	100	93
92	78	80	79
31	52	40	10
53	23	20	33
16	14	4	7
2	10	20	10

Figure 3-11 Cluster CCF is bootstrapped 100 times from the mutation CCF through sampling with replacement.

3.3.4 Bayesian Information Criterion

The bootstrapped RSS values are converted into Bayesian information criterion for each sample and then used for model comparison (Schwarz 1978). BIC is well suited for this application as it is used for model selection under a finite set of models, which consider both the maximised likelihood function and the number of parameters of each model. This results in selecting a model that is the most fitting from the ones compared while punishing overfitting by weighting the number of parameters of the model (Equation 18).

$$BIC = \ln(n)\,k - 2\ln(L) \tag{18}$$

where: *k* is the number of parameters, *L* is the maximized likelihood function and *n* is the number of data points.

In this context, each branch is a possible model. This includes a distinct model for the trunk, corresponding to the null hypothesis, as it models the special case where no subclonal event can be detected (Figure 3-12). All subclonal models have two parameters (CN_{before} and CN_{after}) while the null has only one (since $CN_{before} = CN_{after}$).

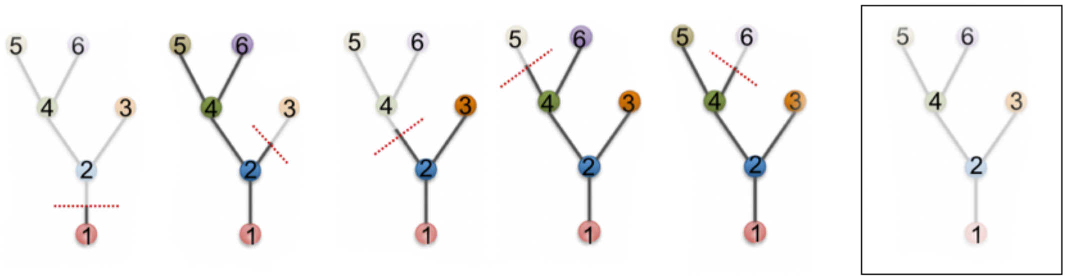


Figure 3-12 MAPSCE includes null hypothesis directly during its model selection.

In this case, the maximized likelihood function is equivalent to RSS. When comparing models, the lowest BIC denotes the most fitting model. This approach provides an objective criterion when choosing the most fitting branch. Because of the bootstrapping, comparing BIC in MAPSCE involves comparing the distributions of BIC for the different branches. It is worth noting that bootstrapping does not apply to the tree trunk as the CCF for the whole tree is, by definition, 100 (Figure 3-13). Since the model for the trunk represents the null hypothesis, we only consider the branches for which at least 95% of the bootstrapped BICs are lower than the null. If no branch fulfils these criteria, the null hypothesis is not rejected and MAPSCE concludes that no subclonal event can be inferred.

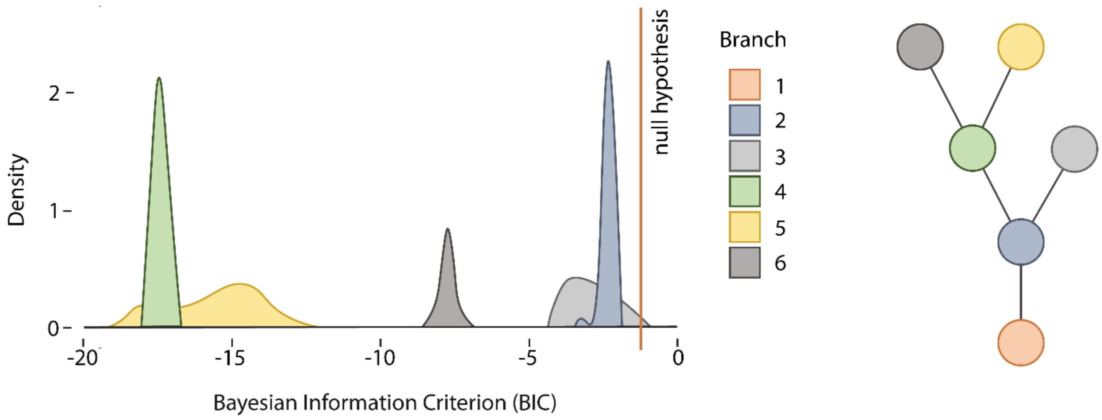


Figure 3-13 Bootstrapped BIC distributions are compared to each other, and to the null hypothesis BIC.

3.3.5 Bayes factor comparison

After considering the distribution of the bootstrapped BICs, the mean of the BICs was converted into Bayes Factors:

$$BF = e^{\left(\frac{BIC_2 - BIC_1}{2}\right)} \tag{19}$$

where BF is the Bayes factor for the chosen branch, BIC_2 is the BIC value of the next best branch and BIC_1 is the BIC value of the top branch. Bayes factors of every branch of a particular allele were then compared to the BF of the top branch to determine the strength of evidence for the most fitting one according to this grading (Kass and Raftery 1995):

Table 3-1 Grading strength of BF difference between the lowest and the second-lowest model (Kass and Raftery 1995).

Evidence for most fitting model	ΔBF to the top branch's BF
positive	0-6
strong	6-10
very strong	>10

For a secondary model to be considered a good result, the difference between the model's BF and the best model's BF had to be lower than 6 ∆BF. MAPSCE considers both models to be sufficiently good. Lastly, if the top subclonal model's BF is not sufficiently better than the null (∆BF > 6) then the null hypothesis is not rejected. This conservative approach ensures that a subclonal event is only called if there is sufficient evidence to reject the null hypothesis. Importantly, the grading for assessing the strength of the evidence for model comparison using BFs was derived in a mathematical model without the biological context provided in this analysis. Other phylogenetic tools have previously used BFs for model comparison with a high evidence for the strong support of the model of 20 ∆BF (Drummond and Rambaut 2007). Thus, we decided to only employ a very conservative approach for considering the proximity between two models of 6 ∆BF rather than 10 – 20 ∆BF.

When considering copy number events, rejecting the null can be interpreted as either a clonal CN event or a lack of CN. The interpretation is left to the user as it is dependent on the type of data. For instance, the expectation will be different depending on whether allele-specific or global CN is considered.

3.3.6 Post-mapping filtering

In addition to the statistical filters, MAPSCE also considers the biological relevance of the result. For CN events, MAPSCE considers the difference between CN_{before} and CN_{after}. It requires this difference to be at least 0.4 for a result to be considered subclonal. However, this threshold can be adjusted to cater for different types of data. For copy number specifically, especially allele-specific copy number events, 0.4 has been identified as the most fitting threshold for the event to be considered subclonal. We examined the distribution of the absolute differences between the CN_{before} and CN_{after} inferred by MAPSCE for simulated subclonal CN events with 0% noise (described later in Chapter 4) based on trees and mutational CCFs from the NSCLC TRACERx 100 data (Jamal-Hanjani et al. 2017). The 0.4 value was chosen to prevent overcalling subclonal events while still providing a high detection rate (Figure 3-14).

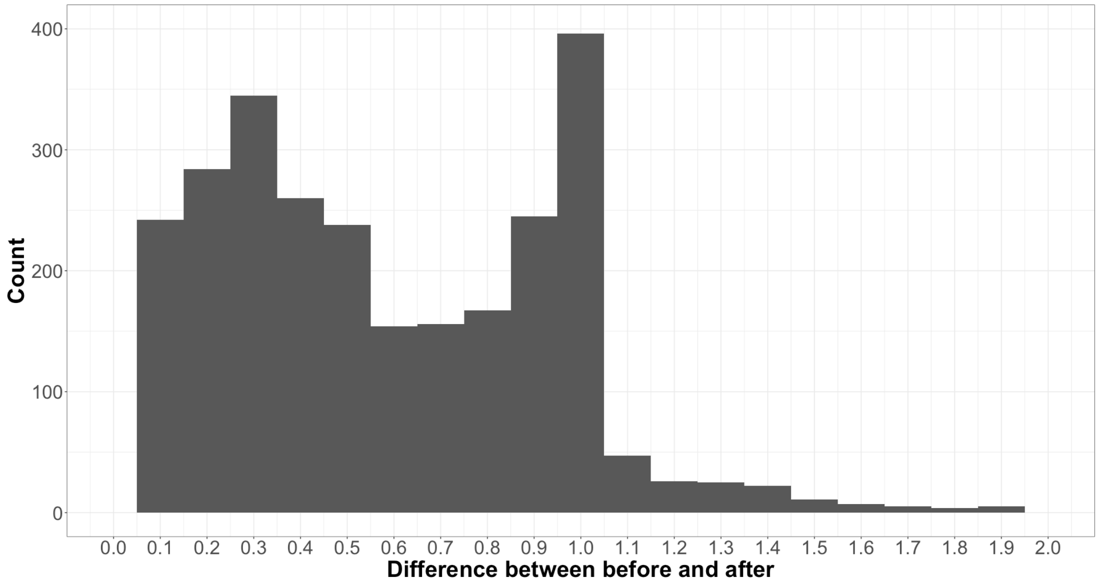
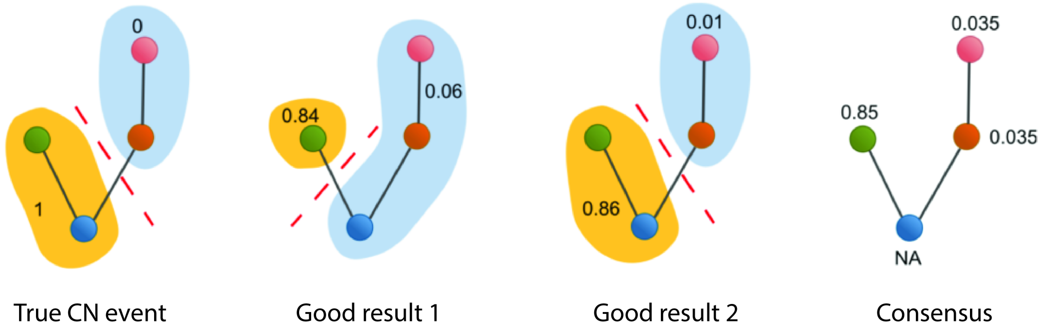


Figure 3-14 Absolute difference between CN_{before} and CN_{after} for the simulated subclonal CN events inferred by MAPSCE.

3.3.7 Consensus mapping

In some cases, there is more than one good result that is statistically significant and significantly better than the null hypothesis, meaning 95% of bootstrapped BICs are lower than the null hypothesis' BIC, the Bayes Factor is within 6 ∆BF to the top model's BF, the BF exceeds null hypothesis' BF by 6 ∆BF, and the difference between CN_{before} and CN_{after} is at least 0.4 (in default settings for subclonal copy number events).



True CN event

Good result 1

Good result 2

Consensus

Figure 3-15 Consensus mapping. Dashed lines denote branches which branch test identified as the best answers. On the left side the simulated event, in the middle the two good results and on the right side the consensus mapping for the tree.

In these cases, each model produces an inferred copy number state for each clone (Figure 3-15). MAPSCE consolidates the results by calculating the average of the inferred copy number states for each clone. This is done for all clones where the different models agree. In practical terms, MAPSCE calculates the mean of the inferred CN states for each clone and calculates the difference between individual inferred CN states and the mean. If for a particular clone, any of these differences are larger than a set threshold (by default 0.1), the clone is left without a consensus CN state. If, following consensus mapping, fewer than two clones have a consensus state, the top-scoring solution is selected instead. The equations for consensus mapping testing as well as calculations for an example case were:

$$Result_1: CN_{before} = 0.84 \;\&\; CN_{after} = 0.06$$

$$Result_2: CN_{before} = 0.86 \;\&\; CN_{after} = 0.01$$

$$Consensus\ test_{before.1}: \left|\frac{CN_{before.1} + CN_{before.2}}{n} - CN_{before.1}\right| < consensus\ threshold$$

$$Consensus\ test_{before.2}: \left|\frac{CN_{before.1} + CN_{before.2}}{n} - CN_{before.2}\right| < consensus\ threshold$$

$$Consensus\ test_{after.1}: \left|\frac{CN_{after.1} + CN_{after.2}}{n} - CN_{after.1}\right| < consensus\ threshol$$

$$Consensus\ test_{after.2}\colon \left|\frac{CN_{after.1}+CN_{after.2}}{n}-CN_{after.2}\right| < consensus\ threshold$$

$$Consensus\ test_{before.1}\colon \left|\frac{0.84+0.86}{2}-0.84\right| \leq 0.1$$

$$Consensus\ test_{before.2}\colon \left|\frac{0.84+0.86}{2}-0.86\right| \leq 0.1$$

$$Consensus\ test_{after.1}\colon \left|\frac{0.06+0.01}{2}-0.06\right| \leq 0.1$$

$$Consensus\ test_{after.2}\colon \left|\frac{0.06+0.01}{2}-0.01\right| \leq 0.1$$

$$Consensus\ mapping_{before}\colon CN_{before}=\frac{0.84+0.86}{2}=0.85$$

$$Consensus\ mapping_{after}\colon CN_{after}=\frac{0.06+0.01}{2}=0.035 \qquad (20)$$

3.3.8 Possible modifications to the default algorithm

3.3.8.1 Mapsce2r

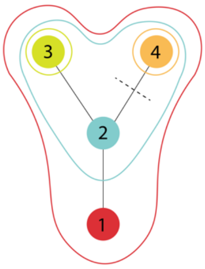
Following testing with the simulated data (described later in Chapter 4), we identified the number of regions as a limitation of the tool's performance. Specifically, for cases, where only two regions are available, MAPSCE automatically performs all of the aforementioned calculations without bootstrapping. Thus, rather than comparing the bootstrapped BICs, only single BIC values for one branch test per branch are compared. This feature, mapsce2r, allows for improved mapping accuracy for cases with two regions only, since no additional noise from the bootstrapping is added to the data. For more than two regions, bootstrapping proved to be a feature that considerably improved the mapping accuracy. These results will be covered in Chapter 4.

Furthermore, testing the tool on simulated data also showed that in all cases, but especially the ones with two regions, setting specific constraints on the

CN_{before} and CN_{after} for the data based on the expected type of events showed great improvement to the mapping accuracy (Chapter 4).

3.3.8.2 Comparing cluster CCF and clone CCF

As described previously, cluster CCF represents the proportion of cells harbouring a particular set of mutations. This is the proportion of cells in a particular lineage. In the initial LOHHLA mapping approach, cluster CCF was converted into clone CCF which was then used for the quadratic programming-based branch test (McGranahan et al. 2017). Clone CCF is the proportion of cells with the same genotype, unique for that particular clone. Figure 3-16 shows the example tree, copy number data, cluster, and clone CCF with the example calculations of the CN states before and after the branch. In an ideal scenario, the calculations using cluster CCF and clone CCF should be identical. However, converting cluster CCFs to clone CCFs is fallible due to the estimated clone CCFs oftentimes not adding up to 100%. This makes using cluster CCF the safer option. Additionally, there is circularity in converting cluster CCFs to clone CCFs only to then add up the clone CCFs for clones before and after the tested branch.



Copy number			
R1	R2	R3	R4
0.4	0.3	0.5	0.8

Cluster CCF

	Region 1	Region 2	Region 3	Region 4
1	100	100	100	100
2	80	90	95	85
3	40	20	35	40
4	30	50	30	15

R1: $0.4 = 70CN_{before} + 20CN_{after}$
R2: $0.3 = 50CN_{before} + 50CN_{after}$
R3: $0.5 = 70CN_{before} + 30CN_{after}$
R4: $0.8 = 85CN_{before} + 15CN_{after}$
→ $CN_{before} = 0.73$ & $CN_{after} = 0$

Clone CCF

	Region 1	Region 2	Region 3	Region 4
1	20	10	5	15
2	10	20	30	30
3	40	20	35	40
4	30	50	30	15

R1: $0.4 = 20CN_{before} + 10CN_{before} + 40CN_{before} + 20CN_{after}$
R2: $0.3 = 10CN_{before} + 20CN_{before} + 20CN_{before} + 50CN_{after}$
R3: $0.5 = 5CN_{before} + 30CN_{before} + 35CN_{before} + 30CN_{after}$
R4: $0.8 = 15CN_{before} + 30CN_{before} + 40CN_{before} + 15CN_{after}$
$CN_{before} = 0.73$ & $CN_{after} = 0$

Figure 3-16 Example tree, copy number data and cluster CCF, and clone CCF with their respective calculations of the CN$_{before}$ and CN$_{after}$.

Instead of converting cluster CCFs to clone CCFs, cluster CCF can be directly used in quadratic programming. Assuming that the trunk always denotes 100% CCF, the branch tested in the branch test splits the tree towards the cluster CCF of the lineage after the branch, and the cluster CCF of the lineage before the branch. Thus, conversion to individual clone CCFs of every clone is avoided. Consequently, branch testing with just cluster CCF lead to the more efficient calculation.

These two approaches were tested against each other on the HLA LOH events in the TRACERx 100 NSCLC dataset. The accuracy of the two approaches was dependent on the RSS between the inferred CN before and after the branch using quadratic programming and the closest integer. Equation 21 describes the RSS calculation for this test. As established before, the better the result was denoted by the lower RSS. Let $[x]$ mean the integer closest to x (rounding up for half – integer values):

$$RSS = \left|CN_{before} - \left[CN_{before}\right]\right|^2 + \left|CN_{after} - \left[CN_{after}\right]\right|^2 \qquad (21)$$

where RSS is the sum of squares of residuals, CN_{before} is the copy number state before the branch and CN_{after} is the copy number state after the branch.

The RSS for cluster CCFs was slightly lower than the RSS for clone CCFs (39.97 for cluster CCFs and 40.92 for clone CCFs) when looking at the CN states before and after for all the branches of every patient of TRACERx 100 (Jamal-Hanjani et al. 2017). When comparing the good branches only, the RSS for cluster CCFs was still slightly lower than the RSS for clone CCFs (0.0210 for cluster CCFs and 0.0215 for clone CCFs). Thus, while the tool allows the user to choose between using the cluster CCFs or clone CCFs in quadratic programming, the cluster CCF is the default option.

3.3.9 Using MAPSCE

3.3.9.1 Inputs

MAPSCE requires multi-sample sequencing data. As input, four types of information need to be provided, which include:

- Observed data (i.e. regional copy number value)
- Mutation CCF (required for bootstrapping specifically)
- Cluster CCF
- Tumour evolutionary tree

The tool provides two different example inputs, including copy number data, mutation CCF, cluster CCF, and tree for a case with two regions only and for another case with more than two regions.

3.3.9.1.1 Observed data

Observed data need to be provided as a numerical vector containing values for each region (minimum two regions). For example, for copy number, each respective value in the numerical vector would be the observed copy number value for a particular region.

3.3.9.1.2 Mutation CCF

Mutation CCF needs to be provided as a data frame, where each row represents a particular mutation, while the columns denote the corresponding CCF of that mutation in each region. The CCF values in the mutation CCF should be in decimals, rather than a percentage. Two additional columns named “PycloneCluster” and “CleanCluster” need to be provided. These columns are included in the output of the upstream analysis of the NSCLC TRACERx 100 methodology (Jamal-Hanjani et al. 2017) involving PyClone (Roth et al. 2014). “PycloneCluster” is the assigned cluster of the mutation, while the “CleanCluster” denotes whether the cluster passes through two filters. The first filter checks the size of the cluster and whether it has at least 5 mutations. The second filter tests whether the cluster is copy number driven, i.e. whether the cluster is absent because of a copy number loss rather than the fact the mutations were never present. “CleanCluster” values need to be either 1 or 0, where 1 shows that the cluster has passed the aforementioned filters to be considered for the analysis, while 0 suggests the cluster and the mutations should be disregarded. This format is based on the output of PyClone (Roth et al. 2014) and reflects the results produced in NSCLC TRACERx 100 project (Jamal-Hanjani et al. 2017).

3.3.9.1.3 Cluster CCF

While this information can be derived from the mutation CCFs, these are only required for bootstrapping. When no bootstrapping is used, MASPCE uses the cluster CCFs instead of the mutation CCFs. Cluster CCF needs to be a matrix, where each column represents the different regions and the rows represent the different clones. The naming of the rows in the cluster CCF should correspond to the “PycloneCluster”, specifically the ones marked as suitable for the analysis (“CleanCluster” = 1), while the regions in the cluster CCF should be identical to the regions in the mutation CCF data frame. Each respective CCF value in cluster CCF should be a percentage between 0 and 100.

3.3.9.1.4 Tumour evolutionary tree

Finally, MAPSCE requires a tumour evolutionary tree, in the form of a matrix with two columns. In this matrix, the first column denotes the ancestral clone (parent), while the second column denotes the descendant one (child). Thus, every row of this matrix denotes a branch. By definition, the clone present in the first column only is the root node, while all the clones in the second column only are the leaves (tips) of the tree. Every value of the matrix should be a character vector.

3.3.9.2 Other parameters

The tool allows the user to change other parameters, which include:

- the number of bootstraps (by default 100 bootstraps),
- the option to run bootstrapping (by default bootstrapping is used, unless there are only 2 regions in the data),
- the option to use clone CCF for quadratic programming (by default the tool uses cluster CCF for QP),
- the option to set the constraints in the quadratic programming on the before and after states (both by default 0).

Furthermore, the tool can print the raw matrix of results and the mapping duration.

3.3.9.3 Interpretation function

The tool includes a wrapper function for interpretation of the raw results of MAPSCE, called *interpret_mapsce()*. This function provides the interpretation of the mapping results of MAPSCE, integrating the consensus mapping as well as automatically determining the clonality for the user. The input for this function requires the output of MAPSCE and the patient tree used for the output of MAPSCE. Furthermore, this function allows the user to change the minimum difference between the before and after states for the event to be considered subclonal (by default 0.4), the consensus threshold (by default 0.1) and the format of the output (by default a data frame).

3.3.9.4 Outputs

There are two outputs to the tool. The first output is the output of the *MAPSCE()* function, which is the raw mapping result for a particular gene. This output includes

- the branch identifiers denoted by the clone which the branch precedes, i.e. branch is the branch directly before clone 5,
- null hypothesis identifier,
- QP-inferred before and after states,
- number of regions and clones,
- statistics for model selection with a summary of whether the result passes all the statistical filters, including:
    - RSS
    - number of bootstrapped BICs better than the null BIC,
    - the mean BIC,
    - filter for whether the Bayes Factors are within 6 ∆BF to the top one.

The results are sorted by the good solutions at the top, and then by ascending BIC.

The second output is the result of the *interpret_mapsce()* function, which is the interpretation function for the raw mapping results of MAPSCE. This output includes:

- the branch identifier,
- the consensus mapping, containing the consensus states for clones of the tree, where the agreement was found,
- the clonality of the event mapped (subclonal or null/clonal) inferred by MAPSCE,
- the consistency filter, containing information on whether there were at least two good solutions in agreement.

3.4 Discussion

The increased generation of multi-region sequencing data improved the understanding of tumour evolution dynamics (Yan et al. 2019; Gerlinger et al. 2012; Jamal-Hanjani et al. 2017). Increasingly new methods have been developed to analyse the generated data (Van Loo et al. 2010; Roth et al. 2014; Deshwar et al. 2015; Niknafs et al. 2015). However, the majority of the work has focused on single-omics data, and there has been an increasing need for tools designed specifically for the Integration of multi-omics data.

McGranahan et al. (2017) were the first group to present a method to integrate subclonal copy number losses with SNV events in the context of phylogenetic tumour evolution. Their method, LOHHLA, helped identify losses of HLA as distinct events occurring on separate branches of four patients' tumour phylogenetic trees, indicative of parallel evolution. The same HLA alleles were subject to loss on those distinct branches, suggesting that the losses of those HLA alleles were required for subclonal expansion (McGranahan et al. 2017). LOHHLA was a proof of concept study of mapping of subclonal LOH events on a tumour evolutionary tree.

One important distinction between MAPSCE and LOHHLA is their clonality determination (Figure 3-17). As mentioned previously, LOHHLA relies on a heuristic approach based on the observed copy number states in the regional data to determine clonality of the events. LOHHLA then maps the pre-determined subclonal events on a tumour evolutionary tree. Conversely, MAPSCE uses quadratic programming to simultaneously determine the clonality of the events and map them on tumour evolutionary trees (Figure 3-17).

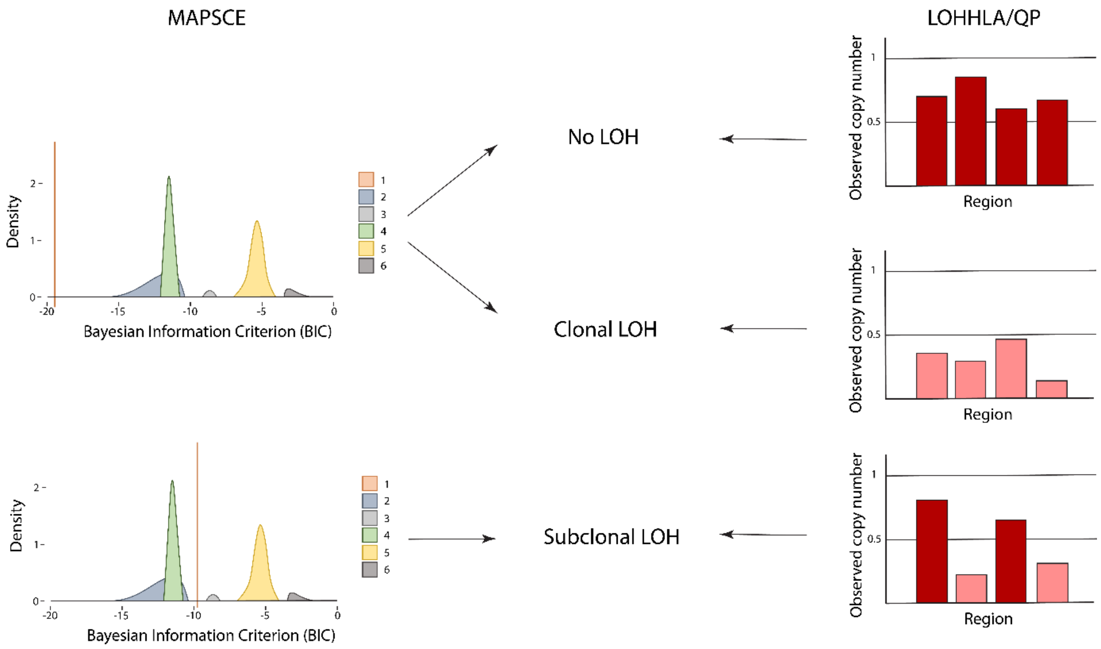


Figure 3-17 Comparison of the clonality determination between MAPSCE and LOHHLA/QP.

This chapter introduced MAPSCE, a tool for mapping subclonal events on SNV-based tumour trees. MAPSCE was specifically designed to handle various types of copy number losses or gains but can accommodate different types of data as well. This chapter detailed the exploration of various methods and algorithms considered during the development of the tool.

3.4.1 Limitations and future work

Deriving CCF from observed copy numbers showed great potential for branch testing, as well as improving the process of clustering of the mutations. Deriving CCF for patients where $CN_{observed}$ is 1 in a region and a fraction (partial loss) in another yielded similar results compared to running quadratic programming, based on the cases where this approach was tested. However, extending this method for handling more complex CN events required estimating the ancestral and derived CN states. This further required the use of quadratic programming. This approach was mainly conceived to provide additional validation for the results obtained with quadratic programming. However, it was deemed inappropriate because of its circularity: QP used $CN_{observed}$ to infer the best fitting branch, and CN_{before} and CN_{after}; these values were then used to derive the CCF_{CN}, necessarily resulting in CCF values close to the ones for the mutations defining the best fitting branch. Further extension of this approach could be useful if those derived CCFs could be used to re-cluster the mutations, leading to more accurate cluster CCFs and trees that already contain information about certain subclonal events. However, this would still require defining original clusters without the CCF_{CN} beforehand to then derive the CCF_{CN} and re-cluster all the $CCFs$. One drawback of such a method is that the CCF_{CN} would be skewed towards the original mutational $CCFs$.

The results of the initial noise simulations were intuitive with smaller clusters being more susceptible to noise. All things considered, one of the confounding factors was identified for further tool development; events simulated on smaller clusters yielded less reliable results. These initial noise simulations were extended to simulate other kinds of CNA for testing the methodology (Chapter 4).

The attempt at using ABC for the branch test was done on simulated data for 10,000 simulations, which resulted in a pseudo-ABC approach that resembled maximum likelihood estimation. Ideally, using ABC in its original formulation would require sampling from the possible solutions for CN_{before} and CN_{after}, which would effectively require a larger number of simulations to sample both these values as well as the different branches. The issue with the full implementation of ABC was how computationally expensive it would be. Ideally, each ABC simulation would include re-clustering of all mutations, however, this would require a very high number of simulations to obtain stable results. Even a simpler sampling approach would require considerable computational resources to run ABC on each patient and every gene in a time-efficient manner.

The filtering process of the results after mapping with a QP-based branch test allowed for accurate detection of the clonality of the events. While the filtering approach limited the number of potential results of MAPSCE, they needed to be refined, as they still were not sufficient to pinpoint the exact branch where the event occurred in certain cases. Despite the filters, the results still included too many potential branches where the LOH event was likely to map without a robust criterion of distinguishing between the solutions. Furthermore, the filtering needed to be extended to also allow for mapping of different copy number changes and other types of events, rather than just CN LOH events. Thus, the statistical filters were further refined as described and then included in the release of the tool on GitHub (v1.0.0). Together with consensus mapping, the filtering approaches allowed for a more accurate determination of clonality and mapping of events.

While both MAPSCE and LOHHLA/QP rely on quadratic programming in their branch test, MAPSCE includes additional statistical features to provide a more robust model selection and a measurement of the goodness of fit of the results. Finally, adding null hypothesis testing directly to the branch test and allowing for either more generic or ad hoc constraints in the quadratic programming compared to LOHHLA allows for the mapping of different data types compared to just HLA LOH events.

The tool, MAPping SubClonal Events, for mapping subclonal events on the tumour evolutionary trees was developed as an R package, available for download on GitHub (https://github.com/MarkTranHS/MAPSCE). In the future, a release for CRAN is also planned. However, that requires additional work to test the software on various operating systems.

In summary, these results present the development and initial testing of the various features included in MAPSCE. The next chapter will describe the testing on simulated datasets to assess the performance of the tool.

Chapter 4 Tool validation using simulated copy number events

4.1 Introduction

The timing of mutational events in cancer provides an insight into tumour evolution, with direct implications for improving patient diagnosis and treatment. Studies have shown how phylogenetic reconstruction of cancers allows for the identification of new subclonal driver events as well as the labelling of previous drivers based on their role in tumour initiation, progression or maintenance (Gerlinger et al. 2012; 2014; Boyle et al. 2021). Nik-Zainal et al. (2012) identified a dominant subclonal lineage comprising more than 50% of tumour cells in every tumour of 21 breast cancer patients using phylogenetic reconstruction. Jamal-Hanjani et al. (2017) have shown that the late subclonal mutations in tumour-suppressor genes occurring after genome doubling predominantly affected only one allele, suggesting that late subclonal events of tumour suppressor genes (TSGs) were often passenger genes, rather than driver events. Miura et al. (2022) have demonstrated how, in lung cancer patients, the influence of smoking-related mutational signatures decreases, while the influence of APOBEC mutational signatures increases during later stages of tumour evolution.

Mapping multiregion copy number data on tumour evolutionary trees improved the understanding of mechanisms underlying tumour evolution in NSCLC (McGranahan et al. 2017) and ovarian cancer (Zhang et al. 2018). In three out of four high-grade serous ovarian cancer patients, samples with subclonal HLA LOH also had the highest epithelial CD8+ TIL densities (Zhang et al. 2018). Subclonal HLA LOH was also linked to subclonal neoantigen depletion in HGSC. McGranahan et al. (2017) further suggested that subclonal HLA LOH is an immune escape mechanism which occurs late in the cancer evolution and facilitates subsequent subclonal expansion. While LOHHLA presented a novel framework for tracking tumour evolution, it lacked a measure of goodness of fit of the results, and its mapping was limited to subclonal HLA

LOH events. Tools for the integration of different types of events extending beyond subclonal LOH in tumour evolutionary context on phylogenetic trees are still lacking. Different approaches were explored and finally, MAPSCE was developed as a new tool to tackle this problem (Chapter 3).

Simulating datasets has always proved to be an useful tool in testing new methodologies in all fields including cancer research (Miller et al. 2014; Roth et al. 2014; Deshwar et al. 2015), allowing for a controlled environment to explore different parameters and optimise the performance of a new algorithm. This chapter presents the testing of MAPSCE on different simulated datasets as well as a comparison of the mapping accuracy of the tool with other existing methodologies. We have utilised the simulated datasets to optimise the performance of various features of the tool, such as bootstrapping and the incorporation of constraints for patients with two regions and more than two regions sequenced. The simulated datasets also allowed us to explore the noise present in the trees provided in the NSCLC TRACERx 100 dataset (Jamal-Hanjani et al. 2017).

In this chapter, the simulated events for patients with two regions sequenced are henceforth referred to as “events with two regions”. Analogically, simulated events for patients with more than two regions sequenced are henceforth referred to as “events with more than two regions”. As stated in the previous chapter, we identified the number of regions as one of the factors affecting the accuracy of the results. Having only two regions sequenced leads to an issue of overfitting, as we provide quadratic programming with a system of two equations with two unknowns. As a result, when evaluating the performance of the tools, we categorised the results based on clonality and number of regions.

Additionally, we measured the performance of the tools based on their recall, the percentage of simulated events that were correctly identified, and their precision, the percentage of the identified results that were correct. The recall and precision can be measured at different levels. In this chapter, we examined either “perfect performance” (or “perfect recall” and “perfect precision"), where the clonality, branch and CN states have been correctly inferred, or the performance for less stringent criteria as well. These criteria were nested within each other. The "clonality level" involved correctly determining an event as clonal or subclonal, as appropriate. The "branch level" for subclonal events referred to correctly identifying the branch where the subclonal event occurred or the mapping of the event on the trunk for clonal events. The "CN level" denoted correctly predicting the clonality, the branch and the CN states (within a tolerance of 0.3). The majority of the results in this chapter were tested for perfect performance (recall or precision) unless otherwise specified. Finally, the simulated events in this chapter included the addition of noise. For each event, the noise value denoted the maximum level of noise that could have been added to a particular simulated CN event. As such, a 15% noise "range" or "level" henceforth referred to ±15% noise, i.e. up to 15% added or subtracted to the CN (a CN range of 0.85 – 1.15 CN in 15% noise range for CN = 1).

4.2 Results

4.2.1 Testing bootstrapping and constraints

MAPSCE allows the user to choose whether to use bootstrapping and whether to set custom constraints on the inferred results. We examined the effects of the constraints and bootstrapping on MAPSCE's recall to determine suitable default values for these parameters (Figure 4-1). When using bootstrapping, MAPSCE was run with 100 bootstrap samples. When using constraints, these were CN_{before} higher or equal to 0.5 and CN_{after} higher or equal to 0. Otherwise, both CN_{before} and CN_{after} were forced to be non-negative.

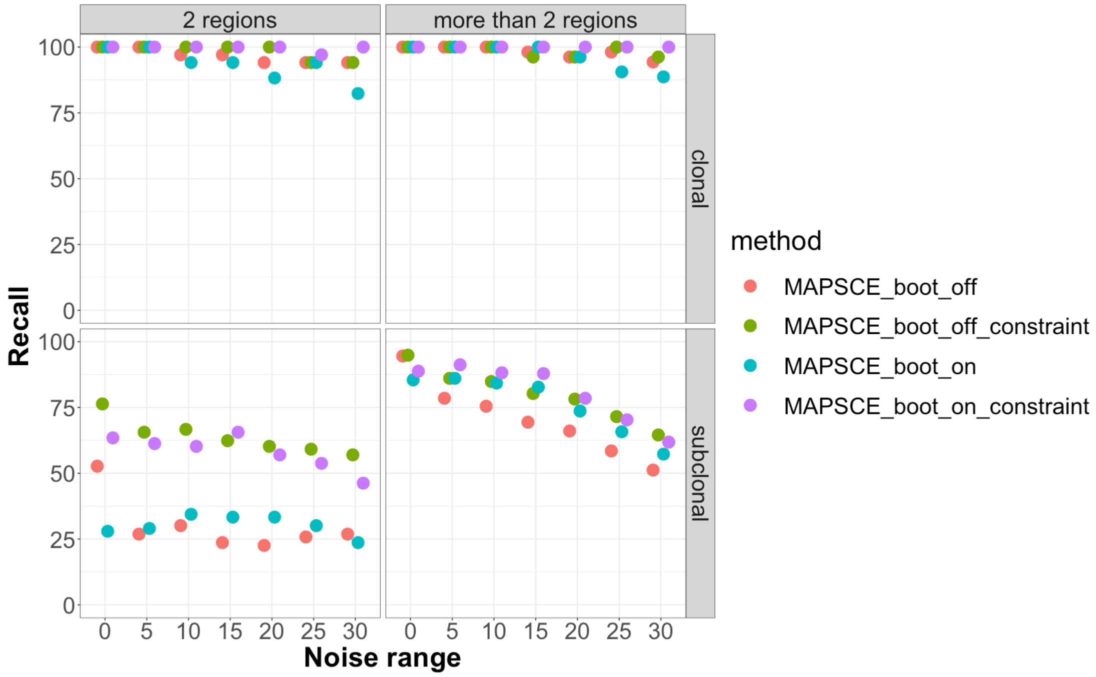


Figure 4-1 Testing of MAPSCE's features (boot_off – without bootstrapping, boot_on – with bootstrapping, constraint – with constraints, no constraints if not specified), and their respective effects on recall in the simulated LOH events dataset with different noise ranges (0 – 30%). Top: clonal; bottom: subclonal; left: two regions; right: more than two regions.

The recall generally decreased with increasing noise levels, with subclonal events having a more pronounced decrease than clonal events.

For clonal events, MAPSCE maintained a high recall above 75% for every noise range, regardless of the number of regions or the different combinations of parameters. MAPSCE with bootstrapping in events with two regions performed the worst, with its recall dropping below 90% in 20% and 30% noise levels.

It was more challenging for MAPSCE to correctly map subclonal events with two regions sequenced. In those cases, the addition of a constraint improved the recall by ca. 25 – 40%. Without these constraints, quadratic programming employed in MAPSCE could easily generate mathematically valid, albeit biologically non-sensical results where the inferred event is a gain from CN=0 to CN=1 on the sister lineage rather than a loss on the correct lineage. This was especially prone to happen in cases with two regions where the cluster CCF of the root node was small in all regions (Figure 4-2). The interpretation function in MAPSCE already filters out these non-sensical results. However, providing quadratic programming with appropriate constraints results in a higher number of results appropriate for the biological context of the dataset.

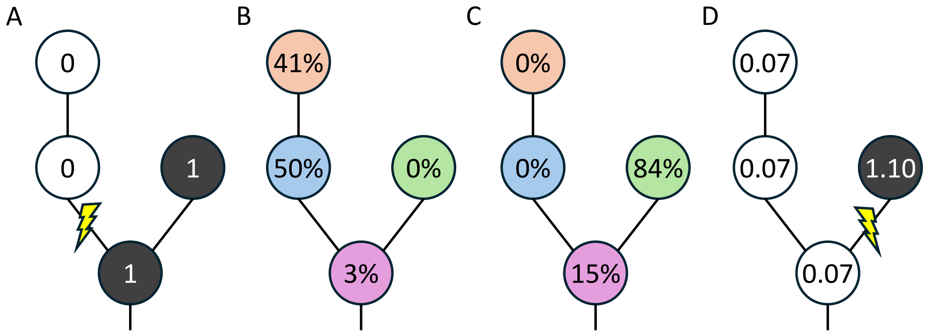


Figure 4-2 Example of common error by MAPSCE when no constraint was provided. Instead of inferring the correct LOH event (A), MAPSCE inferred a loss in the sister branch when up to 10% noise is added (D). B and C show the clone CCF in the two regions for this patient (CRUK0010). In this case, the clone CCF for the root node was small in both regions (3% and 15%) which prevented the precise mapping of the event despite the result being correct for all but the root node. Figure produced by Javier Herrero.

The advantage of using bootstrap for events with two regions was less clear. Bootstrapping improved the clonality determination, however it also decreased the likelihood of identifying the correct branch (Figure 4-3). However, this affected mainly the smaller clusters (0 – 40%). For lineages with a CCF larger than 60%, the mapping recall was noticeably better than on smaller lineages (Figure 4-3).

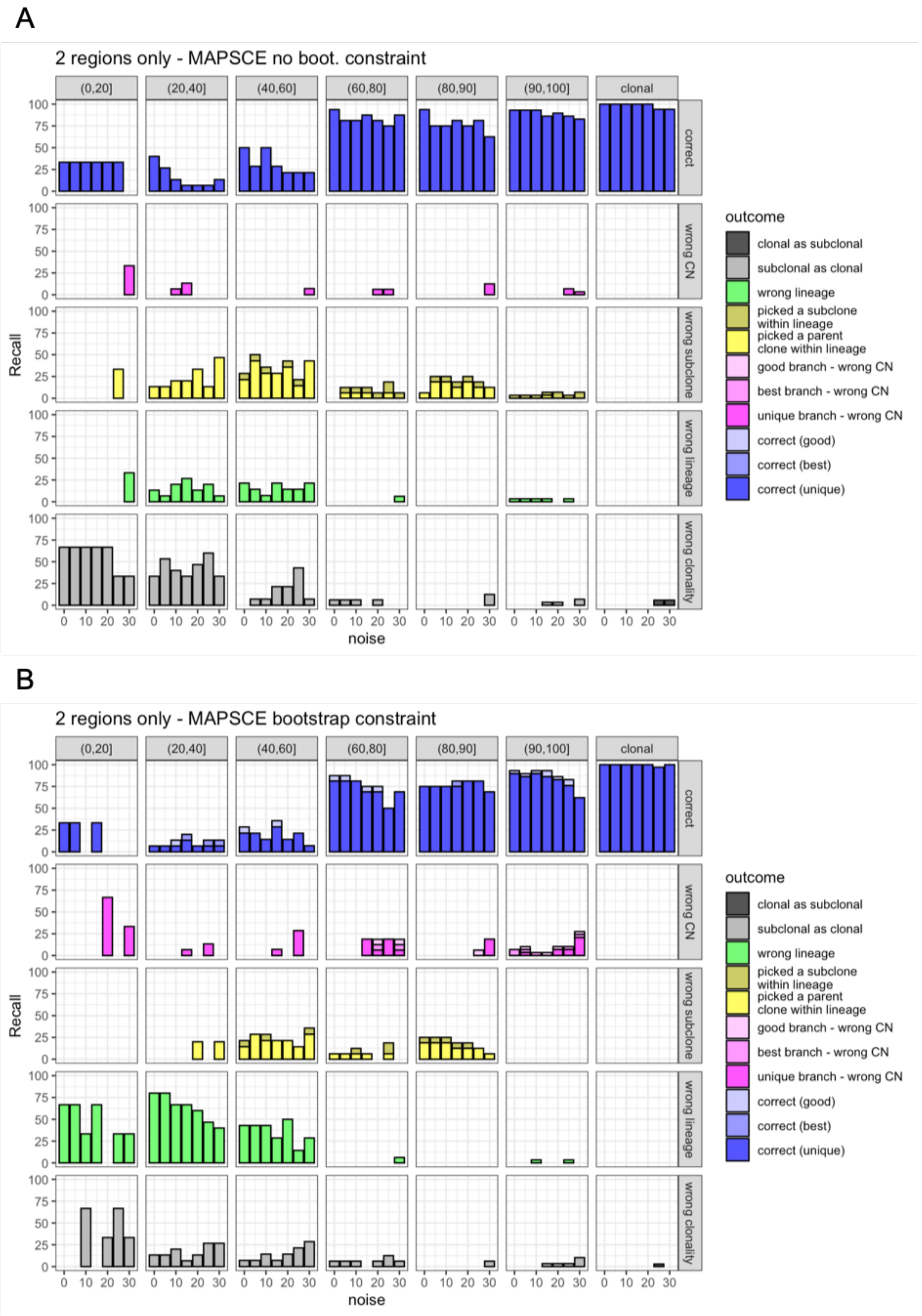


Figure 4-3 Insight into various outcomes of MAPSCE with constraints for events with two regions with (bootstrap) and without (no boot.) bootstrapping in different noise ranges (0 – 30%). The cluster sizes (0 – 100%) relate to subclonal events, while the right category shows clonal events. Bootstrapping improved clonality determination, however led to

a lower likelihood of identifying the correct branch. Figure produced by Javier Herrero.

For subclonal events with more than two regions, MAPSCE's recall declined considerably with increasing noise range (ca. 5-10% decline for every 5% noise increase). However, the tool still maintained above 50% mapping accuracy across all noise ranges (Figure 4-2). For subclonal events with more than two regions, the addition of bootstrapping improved the recall by ca. 10-15% in every noise level. The inclusion of constraints still generally improved the results in simulated subclonal events with more than two regions, but the improvement was less pronounced (ca. 5-10%) compared to the addition of constraints in subclonal events with two regions.

These results suggest that including constraints to restrict the solution space was the largest contributor to MAPSCE's improved performance for subclonal events with two regions. In these cases, bootstrapping improved the clonality determination, while reducing the likelihood of correctly identifying the correct branch. Conversely, for subclonal events with more than two regions, the addition of bootstrapping was the largest contributor to MAPSCE's improved recall. The advantages of including constraints were less prominent in this case, however including both the bootstrapping and the constraints led to the largest improvement in the performance for subclonal events with more than two regions.

4.2.2 Comparing MAPSCE to LOHHLA/QP on a dataset with simulated loss of heterozygosity events

We compared MAPSCE to LOHHLA/QP using the simulated loss of heterozygosity events, as LOHHLA's mapping approach was specifically designed to address these types of events. We ran quadratic programming in MAPSCE with the same constraint as in LOHHLA, namely the inferred ancestral copy number state (CN_{before}) had to be at least 0.5. We employed the default parameters of MAPSCE for bootstrapping. These included bootstrapping for events with more than two regions but none for events with two regions. In this analysis, we looked at the aforementioned different levels of correctness, the clonality, the branch and the CN state.

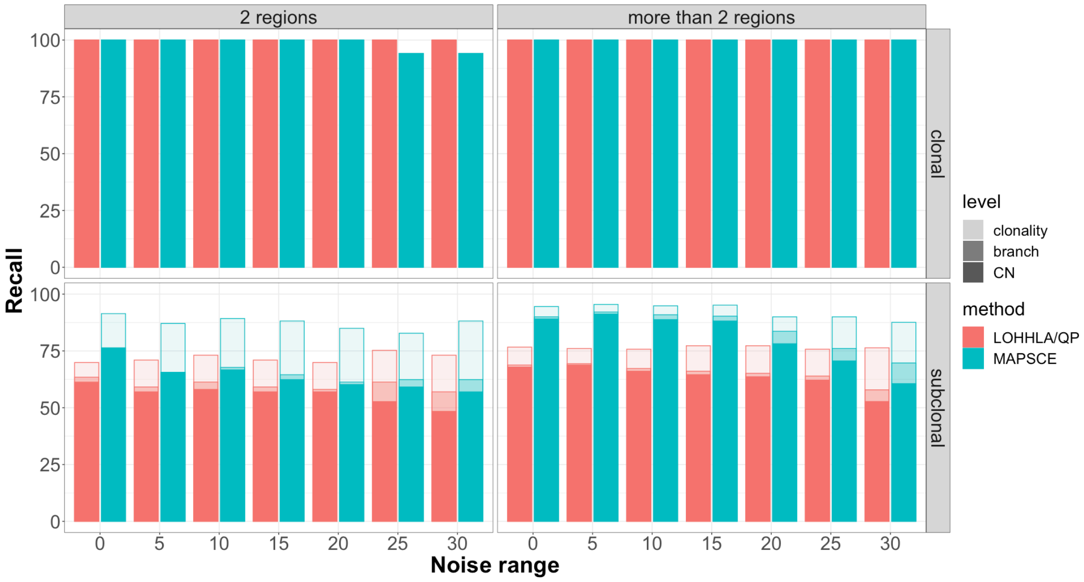


Figure 4-4 Recall of LOHHLA/QP and MAPSCE on simulated LOH events with different noise levels (0 – 30%). Top: clonal results; bottom: subclonal results; left: cases with two regions only; and right: cases with more than two regions. Different intensity levels show the various criteria of correctness considered. Clonality: correctly identified as a clonal or subclonal event; branch: for subclonal events, correctly identified the branch where the subclonal event occurred; CN: correctly predicted the clonality, the branch and the CN states (within a tolerance of 0.3).

Both tools maintained high recall above 95% in identifying all clonal events regardless of the number of regions (Figure 4-4). Only MAPSCE's performance degraded in clonal events with two regions to ca. 95% when the noise ranges reached 25% – 30%.

MAPSCE consistently outperformed LOHHLA/QP in terms of recall for subclonal events, across all noise ranges. The advantage of MAPSCE's mapping accuracy was more evident in subclonal events with more than two regions (difference of ca. 3 – 15% depending on noise) compared to subclonal events with two regions (difference of ca. 8 – 23% depending on noise). MAPSCE maintained a recall of over 75% with noise levels up to 20% in subclonal events with more than two regions. However, as the noise levels increased, MAPSCE's improved recall over LOHHLA/QP decreased, suggesting MAPSCE's lower tolerance for higher noise levels (25 – 30%).

MAPSCE was considerably better than LOHHLA/QP at correctly identifying subclonal events, regardless of the number of regions (Figure 4-4). Even in the presence of large noise, MAPSCE still remained capable of correctly discriminating clonal from subclonal events, although the noise affected its ability to correctly identify the branch affected by the LOH event.

As previously mentioned, increasing noise levels led to a higher decrease in MAPSCE's performance compared to LOHHLA/QP's. The increasing noise levels mainly affected the smaller, rather than the larger branches. Consequently, we hypothesised that LOHHLA/QP was a more conservative tool, which did not consider smaller clones and mainly mapped events on the larger clusters. Conversely, bootstrapping in MAPSCE led to increased sensitivity for subclonal events simulated on smaller clusters, at the cost of lower tolerance of increasing noise levels.

We tested this hypothesis by comparing the recall and precision of both tools in different cluster sizes. For this analysis, we considered the maximum size of a cluster among all regions as the cluster size, i.e. the size of the cluster in the region where it is the largest.

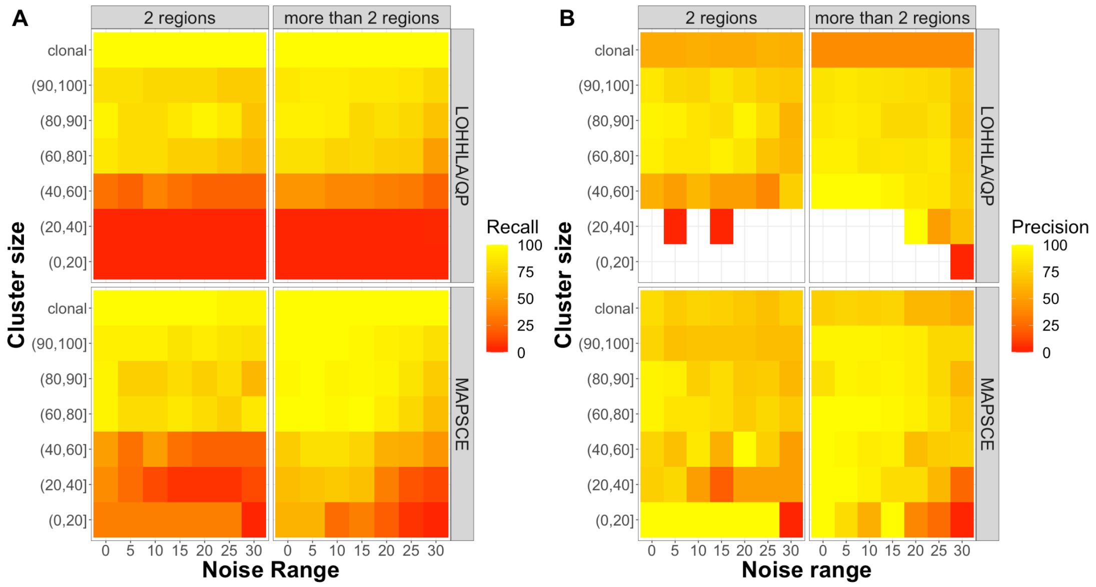


Figure 4-5 Recall (A) and precision (B) of LOHHLA/QP and MAPSCE in different cluster sizes and clonality on simulated LOH events with different noise levels (0 – 30%). Top: LOHHLA/QP; bottom: MAPSCE; left: two regions only; right: more than two regions. Yellow: high recall or precision; red: low recall or precision. The cluster sizes (0 – 100%) relate to subclonal events, while the top category shows clonal events.

In general, LOH events simulated on small (0 – 40%) and medium clusters (40 – 60%) were more challenging to map correctly due to a lower signal compared to the large clusters (60 – 100%). As a result, both tools struggled with clusters of smaller size (Figure 4-5).

LOHHLA/QP was especially prone to mistakes when mapping events simulated on smaller cluster sizes (recall of ca. 0 – 5%) (Figure 4-5A). Additionally, LOHHLA/QP's precision in the smaller clusters showed missing data since LOHHLA/QP mostly did not map events on the small clusters (Figure 4-5B). LOHHLA/QP maintained a high recall regardless of the number of regions for large clusters (ca. 50 – 94%) and for clonal events (100%). However, LOHHLA/QP's precision for clonal events was considerably lower (ca. 54 – 59% for two regions, and ca. 39 – 41% for more than two regions).

As mentioned previously, MAPSCE's recall and precision declined more with increasing noise levels, especially above 20% noise. MAPSCE generally outperformed LOHHLA/QP in terms of their recall and precision for subclonal events regardless of the cluster size. The advantage of MAPSCE's performance over LOHHLA/QP's was particularly noticeable in small and medium clusters, especially in their precision. MAPSCE's precision (up to 20% noise: precision of ca. 20 – 100% for 2 regions, and ca. 37 – 100% for more than 2 regions) was higher than its recall (up to 20% noise: recall of ca. 6 – 40% for 2 regions, and ca. 20 – 70% for more than 2 regions) in small clusters. Importantly, MAPSCE's precision was consistently higher than LOHHLA/QP's precision in clonal events regardless of the number of regions and noise levels.

These results suggest that LOHHLA/QP mapped events conservatively, overcalling clonal events and mostly mapping subclonal events on the medium or large branches. For LOHHLA/QP to map events on small clusters, it required the observed CN value to be less than 0.5 in at least one of the regions. The few cases where LOHHLA/QP did map an event on a small cluster could be attributed to the effect of the added noise, such that the observed CN in a particular region dipped below 0.5 even for a small cluster. This led to an increased recall in large clusters and clonal events, at the cost of a decreased recall in small and medium clusters, and a decreased precision overall. These results also explain LOHHLA/QP's high tolerance to increasing noise levels, as the tool ignored the smaller events that were mainly affected by the added noise. Conversely, MAPSCE was more sensitive in detecting small and medium clusters, at the cost of lower tolerance to increasing noise levels (above 10% noise for two regions and above 20% for more than two regions).

4.2.3 MAPSCE's performance beyond LOH events

We tested MAPSCE on other types of simulated copy number events, including amplifications (two to three copies), duplications (two to four copies), homozygous losses (two to zero copies) and heterozygous losses (two copies to one copy). We used two different sets of constraints for running MAPSCE on these datasets. These constraints were chosen depending on the type of simulated copy number events. For both heterozygous and homozygous loss events, we used $CN_{before} \geq 1.5$, while for the amplifications and duplications, we used $CN_{after} \geq 2.5$.

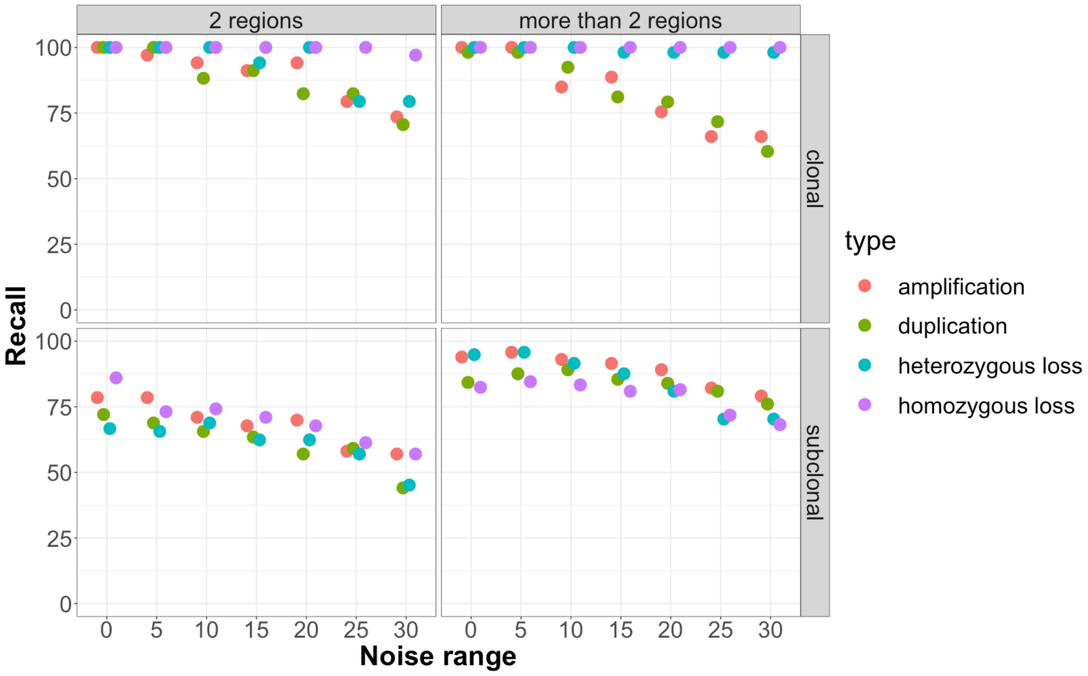


Figure 4-6 Recall of MAPSCE on the dataset with other types of copy number events simulated with different noise levels (0 – 30%). Top: clonal results; bottom: subclonal results; left: cases with two regions only; and right: cases with more than two regions.

The tool maintained a high recall of over 75% for clonal results up to 25% noise for events with two regions specifically (Figure 4-6). For clonal events with more than two regions, MAPSCE maintained over 75% recall in all four datasets only up to 15% noise.

Interestingly, MAPSCE noted a higher mapping accuracy for clonal homozygous losses compared to the other clonal copy number events for two regions. This could be attributed to the fact that homozygous losses were easier to distinguish even in the presence of noise due to the larger drop from $CN_{before} = 2$ down to $CN_{after} = 0$ compared to heterozygous losses and amplifications. The difference between the tool's performance between homozygous losses and duplications could be explained by more appropriate constraints on CN_{before} applied when running MAPSCE for the loss events. MAPSCE achieved over 95% recall for clonal loss events with more than two regions recall regardless of noise levels. The tool's recall of clonal gain events with more than two regions declined considerably with increasing noise levels.

The tool's performance for subclonal events was worse than for clonal events. MAPSCE performed better when mapping subclonal events with more than two regions compared to subclonal events with two regions. The tool maintained high recall when mapping subclonal events regardless of the number of regions. As observed when examining the LOH dataset, MAPSCE's performance declined as the noise levels increased, regardless of the number of regions, across all four datasets.

Additionally, we examined MAPSCE's recall and precision in different cluster sizes in all four simulated datasets. As previously stated, in this analysis, we considered the maximum size of a cluster among all regions as the cluster size, i.e. the size of the cluster in the region where it is the largest.

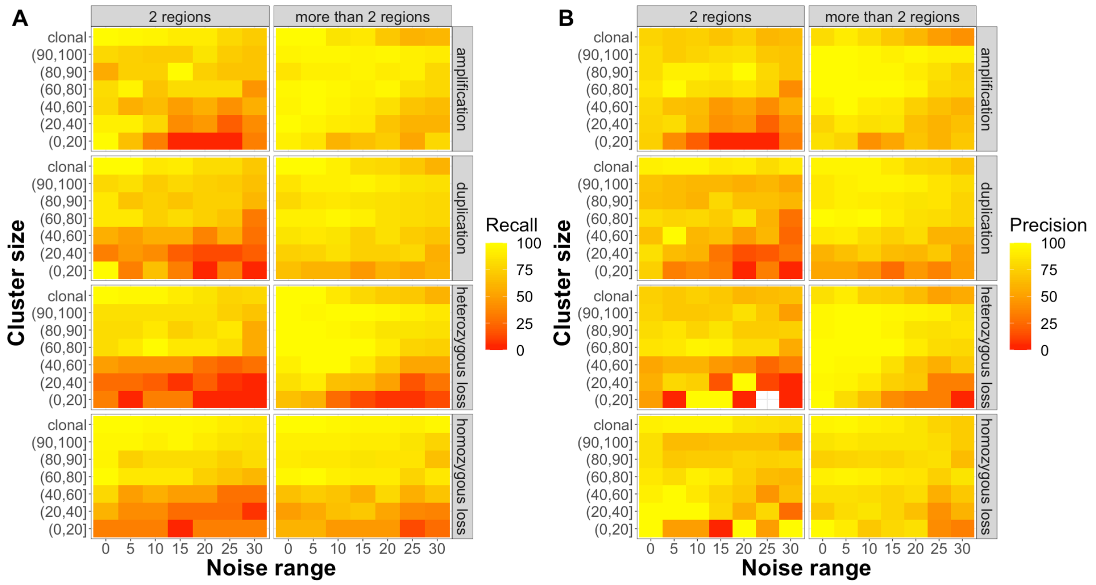


Figure 4-7 Recall (A) and precision (B) of MAPSCE on datasets with different types of copy number events simulated (from top to bottom: amplification, duplication, heterozygous loss, homozygous loss), with different noise levels (0 – 30%). Left: two regions only; right: more than two regions. Yellow: high recall or precision; red: low recall or precision. The cluster sizes (0 – 100%) relate to subclonal events, while the top category shows clonal events.

In general, MAPSCE's performance in other types of CN events was similar across all four simulated types of CN events. The tool's performance did not differ significantly from its performance on simulated LOH events (Figure 4-7). For subclonal events with more than two regions, the tool consistently maintained a high recall and precision regardless of the cluster size of up to 20% noise.

As observed previously, MAPSCE's recall was lower in small and medium clusters compared to the tool's performance in large clusters (Figure 4-7A). This difference in the tool's recall was more evident in events with two regions compared to events with more than two regions. The tool's recall for cases with more than two regions was higher for CN gains than CN losses in smaller clusters, which could be due to the nature of the CNA and the constraints. A gain of CN in the smaller cluster provides a higher signal to be detected.

Additionally, the constraints for copy number gains being $CN_{after} \geq 2.5$ made it less challenging to detect the lineage that was affected by the CNA. When it came to CN losses, the constraints were $CN_{before} \geq 1.5$ and $CN_{after} \geq 0.$ In these cases, the detection of a loss in a small branch was more challenging, since the small clusters already provided a small signal to be detected.

Similarly, MAPSCE's precision was higher in large clusters compared to the small and medium clusters (Figure 4-7B). This difference was again more pronounced in subclonal events with two regions compared to subclonal events with more than two regions. MAPSCE did not map any events in the heterozygous losses dataset for cases with two regions at 0 – 20% CCF and 25% noise, resulting in missing data.

Additionally, we quantified the number of times MAPSCE used consensus mapping to consolidate results. Consensus mapping was utilised in approximately 7 – 10% of cases at 0% noise. This usage steadily increased with increasing noise levels for all four simulated types of CN events, peaking at around 32% for amplifications at 30% noise. This indicates that higher noise levels led to MAPSCE being less precise and classifying more results as mathematically valid. These good results were then merged using consensus mapping.

Taken together, MAPSCE maintained similarly high recall and precision across all four simulated datasets with different types of CNA. The tool performed best when provided with data from patients with more than two regions sequenced. MAPSCE maintained high mapping accuracy up to 20% noise, irrespective of the type of the CNA. This demonstrated that the tool can withstand a lot of noise in the data, even when mapping subclonal copy number gains. MAPSCE's ability to map subclonal copy number gains on SNV-based tumour trees sets it apart from other existing tools.

4.3 Conclusions

4.3.1 Summary of findings

In this chapter, we compared MAPSCE to another approach of a similar purpose, LOHHLA/QP, using simulated LOH events. Additionally, we assessed MAPSCE's performance on other types of simulated copy number alterations, namely amplifications, duplications, heterozygous losses, and homozygous losses. These simulations allowed us to explore MAPSCE's novel ability to map subclonal copy number gains on SNV-based trees, as no directly comparable approach is available to our knowledge. Furthermore, testing the tool on the simulated datasets helped in optimising MAPSCE's default settings. Over the years of the tool's development, the simulations provided a means to evaluate the advantages of the various features included in the tool.

We examined the effects of including constraints and bootstrapping using MAPSCE on simulated LOH events. The tool's performance declined considerably for subclonal events with two regions only. This could be explained by the quadratic programming overfitting results due to being provided with a system of two equations with two unknowns. Additionally, having two regions sequenced only, provided an insufficient amount of data for the tool's optimal performance. The inclusion of constraints considerably improved MAPSCE's recall for subclonal events with two regions. The addition of bootstrapping improved the clonality determination, however decreased the likelihood of mapping the event on the correct lineage.

Conversely, bootstrapping was the largest contributor to MAPSCE's improved recall in subclonal events with more than two regions. While including constraints still led to an increased recall of MAPSCE in subclonal events with more than two regions, this effect was less pronounced than in subclonal events with two regions. Furthermore, incorporating both constraints and bootstrapping for events with more than two regions showed the largest increase in MAPSCE's recall.

Taken together, the default settings in MAPSCE were set to include bootstrapping for data with more than two regions, but not for data with two regions only. The addition of appropriate constraints was always recommended for best performance.

By incorporating various noise ranges from 0 to 30% in the simulated datasets, we aimed to replicate the levels of noise found in real data and gain insight into the resilience of the two approaches to noise. The results indicated that MAPSCE exhibited a more significant decline with increasing noise levels, particularly in events with two regions, compared to LOHHLA/QP. As previously described, LOHHLA/QP was designed to be conservative and ignored events affecting small and medium lineages. The increasing noise levels mainly affected the smaller clusters. As a result, LOHHLA/QP had a higher tolerance to increasing noise levels, at the cost of overcalling clonal events and a lower performance for small and medium clusters. Furthermore, LOHHLA/QP overcalled clonal events, which resulted in a decreased performance when mapping subclonal events. Conversely, MAPSCE's higher recall and precision in the smaller clusters suggest the tool's higher sensitivity to events affecting smaller lineages, at the cost of higher sensitivity to increasing noise levels.

MAPSCE's performance on simulated datasets with other types of copy number alterations was similar to the tool's performance in the simulated LOH dataset. MAPSCE's consistently high recall and precision demonstrated its ability to integrate the copy number data in the context of SNV-based tumour evolutionary trees. As mentioned previously, the tool is prone to mistakes at noise levels above 20%, due to its high sensitivity to mapping events on smaller clusters. Importantly, MAPSCE's ability to accurately map copy number gains is a novel feature, which to our knowledge, has not been attempted by any other existing methodology.

Taken together, the simulated datasets allowed for comparing the performance of MAPSCE with LOHHLA/QP in mapping LOH events. We also assessed MAPSCE's recall and precision on other simulated copy number events. MAPSCE consistently outperformed LOHHLA/QP when mapping simulated subclonal LOH events. The tool also maintained high recall and precision when mapping other simulated copy number events, regardless of clonality. Having assessed the performance of the tool on the simulated data, the next chapter will describe the integration of real multi-omics data using MAPSCE.

4.3.2 Future work

The simulated copy number datasets in this chapter included noise simulations from 0 to 30% noise ranges with uniform distribution. These noise ranges were chosen in the simulated datasets to provide a wide range of potential noise that could be anticipated in real data. However, further work is needed for measuring the noise in the real genomics data. Assessing the extent of noise that is present in the data would provide insight into how distorted the signal in the genomic data is. Additionally, this would allow us to determine how resilient the developed bioinformatics tools need to be to withstand the background noise and still provide accurate results.

The results in this chapter demonstrated MAPSCE's improved recall when provided with appropriate constraints. These constraints could be refined to better provide the tool with the biological context of the dataset. Furthermore, MAPSCE could be improved by including an option for the tool to automatically detect the type of event being mapped. Based on whether the tool identified a CN gain or CN loss, it could also automatically apply the corresponding constraints to quadratic programming. While MAPSCE was primarily tested on simulated copy number events, the tool was designed to be able to integrate any type of data such as gene expression or methylation. The next chapter will provide some results on mapping gene expression changes on the tumour trees along the copy number events. However, further testing of the tool on simulated gene expression changes and methylation events is required.

Chapter 5 Integration of multi-omics data

5.1 Introduction

The rapid advancement in next-generation sequencing techniques has created novel computational challenges in effectively tackling these large-scale NGS data. One major issue of the massive development of NGS data is the increasing scale of the genomic data, which requires improved data integration and interpretation. While tools have been developed to analyse the increasingly complex NGS data, most of them are still single-omics approaches, which do not fully connect the different layers of data (Nik-Zainal, Alexandrov, et al. 2012).

Integrative multi-omics approaches have been crucial for analysing the combined mutational data at different levels to provide a comprehensive understanding of tumour evolution (Silverbush et al. 2019; Schulte-Sasse et al. 2021; Sammut et al. 2022). Silverbush et al. (2019) presented ModulOmics to integrate protein-protein interactions, mutual exclusivity of mutations and CNAs, transcriptional coregulation and RNA co-expression into a single probabilistic model to identify novel cancer driver pathways in breast cancer. Schulte-Sasse et al. (2021) utilised a machine learning approach to combine mutational, CN, DNA methylation, gene expression data and protein-protein interaction networks to predict novel cancer genes. Sammut et al. (2022) also employed machine learning models to provide an integrative, multi-omics approach to predicting therapy response in breast cancer patients. These examples demonstrate how the integration of multi-omics data can provide a more holistic interpretation of the results to identify novel driver genes and cancer pathways for instance.

At its core, MAPSCE allows for the integration of multi-omics analysis into a single framework to provide users with a broader perspective on the data.

What distinguishes MAPSCE from other multi-omics approaches is its ability to integrate data at the tumour clone level. MAPSCE can integrate different types of multi-omics data by using the output from state-of-the-art tools rather than re-interpreting the data.

The previous chapters described the development of the MAPSCE methodology and the testing of the tool's performance on simulated datasets. This chapter will illustrate the potential applications of MAPSCE in integrating multi-region, multi-omics data.

5.2 Results

5.2.1 NSCLC TRACERx 100 HLA LOH

The HLA genes are responsible for presenting the intra-cellular antigens derived from tumour cells to T cell receptors for recognition. There are three main genes (*HLA-A*, *HLA-B*, *HLA-C*) encoding the HLA class I alleles. The HLA locus is highly polymorphic, with thousands of HLA alleles identified for each HLA gene. The polymorphic nature of the HLA region makes accurate copy number analysis of the HLA genes problematic. McGranahan et al. (2017) presented LOHHLA, a tool for determining HLA allele-specific copy numbers from sequencing data. LOHHLA identifies losses of heterozygosity of HLA alleles, determines their clonality and maps the subclonal HLA LOH events on tumour evolutionary trees.

This section aims to validate MAPSCE by testing the tool in the NSCLC TRACERx 100 dataset and comparing it against a previous analysis using LOHHLA (McGranahan et al. 2017). The two methods were compared based on their clonality determination, by classifying each HLA allele as having either no LOH, clonal LOH or subclonal LOH.

5.2.1.1 Overview of the dataset

We first reviewed the NSCLC TRACERx 100 dataset before running both tools to provide a general overview of the dataset (Figure 5-1).

The cluster CCF varied across different sampled regions (Figure 5-1A). For this analysis, the term ‘minimum CCF’ (min CCF) refers to the CCF in the region with the lowest CCF for this cluster, while the ‘maximum CCF’ (max CCF) is the CCF in the region, where the CCF is the largest. Lastly, the ‘average CCF’ (mean CCF) is the average of the CCFs across all regions. As expected, the min and max CCF of the root clusters were ca. 100%, with some outliers of maximum CCF being 95%, due to the noise in the data. The mean CCF for the root clusters was between 94.8% and 100%, with the majority of the mean CCFs falling between 98% and 100% (59 out of 90 root clusters). The CCF of the clusters on the internal branches mostly ranged from 50% to 80% CCF for both minimum and maximum. As expected, the mean CCF of the clusters for the tips of the tree was significantly lower, ranging between 2% and 68.5%. This disparity can be attributed to the vast majority of the tips of the trees being region-specific (254 out of 261 tips). Interestingly, when disregarding the absent regions for the region-specific clusters in the tips of the tree, the average min CCF was comparable to the average max CCF (mean min CCF 57.7%, median min CCF 57%, mean max CCF 60.3%, median max CCF 62%). These results suggest that the tips of the trees comprised mostly region-specific, medium-sized clusters (40-60%), rather than solely small clusters (0-40%). The larger-than-expected size of the region-specific tips of the trees represents the high number of subclonal mutations in the later stages of cancer evolution.

There were between two and four regions sequenced for most patients (77 out of 90, 85.6%) with the mean being 3.2 and the median being 3.0 (Figure 5-1B). As shown in Chapter 4, MAPSCE performed worse in patients with only two regions sequenced, unless additional constraints were provided. In this dataset, 35 out of 90 patients had two regions sequenced (38.9%).

The trees of the patients in this cohort consisted of between 3 to 8 clones on average (72 out of 90 patients, 80%) with a mean of 5.8 and median of 5.0 (Figure 5-1B). The most common patients' trees had 3 to 5 clones (3: 18 patients, 4: 13 patients, 5: 13 patients).

The majority of the patients had clonal whole genome doubling (67 out of 90, which corresponds to 74.4%), with only a few cases being subclonal (3 out of 90, which is 3.3%) and the rest of the patients not having any genome doubling event (20 out of 90, 22.2%) (Figure 5-1C).

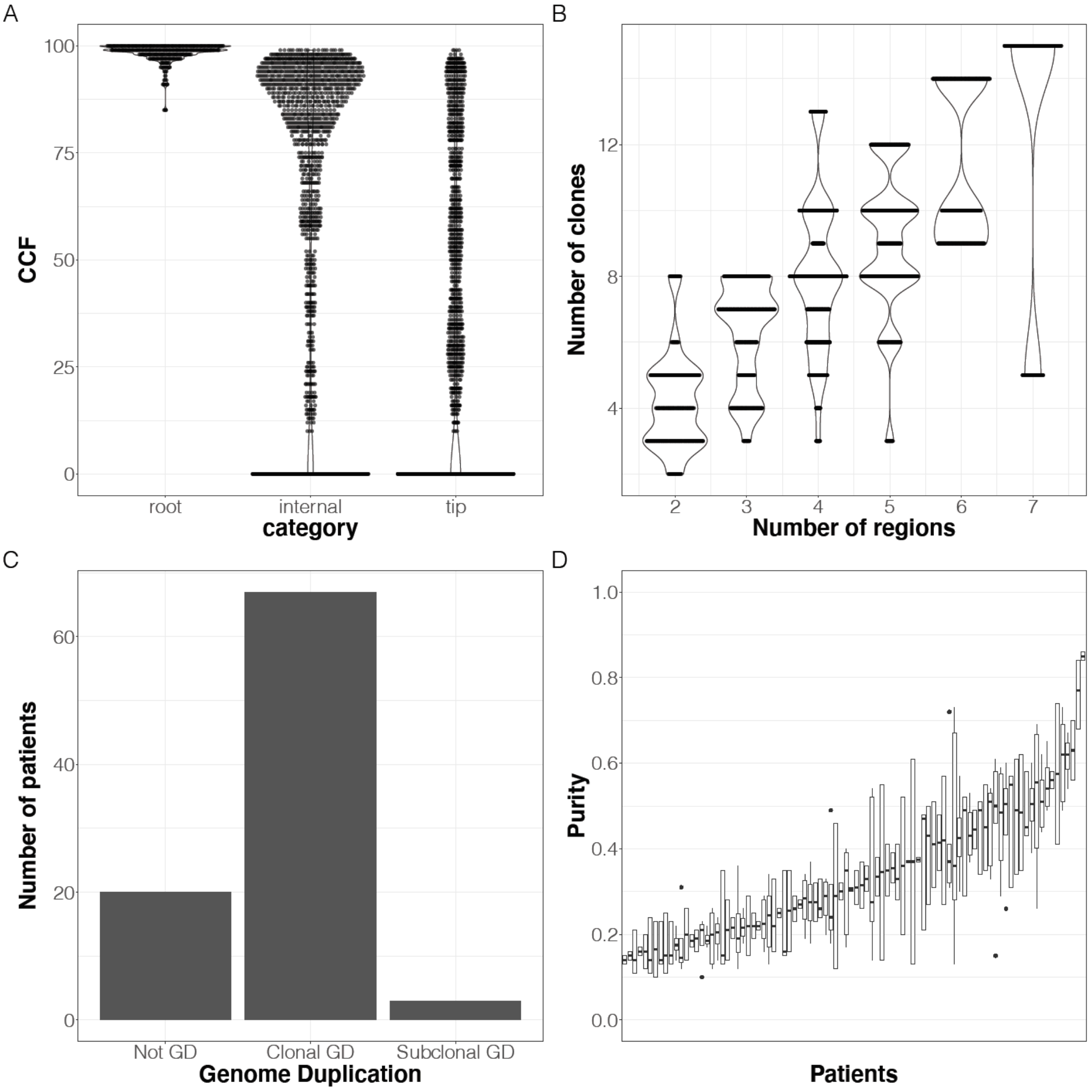


Figure 5-1 Overview of the NSCLC TRACERx 100 dataset. A) CCF distribution for clusters located in different parts of the tree. B) Number of regions and number of clones for each patient of the cohort. C)

Number of patients with different types of whole-genome doubling. D) Purity of the sampled regions for each patient.

Each patient's tumour was sampled in at least two regions, resulting in varying purity for each observed copy number; for each patient, a minimum and a maximum purity can be estimated from all sampled regions (Figure 5-1D). On average, the minimum purity was 25.3%, with a median of 22.0%. The minimum ranged from 10.0% recorded for patients CRUK0004 and CRUK0081, to 84.0% for patient CRUK0084. The maximum purity was 37.0% on average with a median of 42.0%. Values ranged from 15.0% for patient CRUK0064 to 86.0% for patient CRUK0016.

5.2.1.2 Comparison of MAPSCE and LOHHLA/QP in mapping HLA LOH events

In this analysis, out of 100 patients, only 60 patients were considered due to missing patient trees or incomplete copy number data. Among those 60 patients, 288 HLA alleles were analysed, since not every patient was heterozygous for every HLA allele.

The two tools had a 90.3% agreement when classifying HLA alleles as having either no LOH, clonal LOH or subclonal LOH across all 288 alleles (Table 5-1). MAPSCE identified a higher percentage (17.01%) of subclonal HLA LOH events in the dataset compared to LOHHLA/QP (13.89%). For alleles that both tools deemed subclonally lost, the two tools mapped the event on the same branch in 6 out of the 27 cases (22.2%). In total, there were 28 cases where both tools disagreed. In most cases (21 cases), they disagreed on the presence of a LOH event, while in another 7 cases, they disagreed on the clonality of the LOH event (clonal vs subclonal).

Table 5-1 Comparison of MAPSCE and LOHHLA/QP's mapping approach when mapping HLA LOH events in the TRACERx 100 dataset. Numbers denote the number of alleles in each category. "No LOH" indicates alleles where the tool identified no loss of heterozygosity of HLA, "clonal" refers to alleles identified as clonally lost, and "subclonal" are alleles subclonally lost. The bolded numbers in the diagonal show the matching alleles between the two tools.

288 alleles		*LOHHLA/QP*		
		No LOH	*Clonal*	*Subclonal*
MAPSCE	*No LOH*	***205***	*2*	*9*
	Clonal	*0*	***28***	*4*
	Subclonal	*10*	*3*	***27***

When calling subclonal events, MAPSCE's good results, those that satisfied the tool's statistical filters, showed a mean RSS of 0.315, and a median of 0.034. Conversely, when looking at the top result for the subclonal events mapped by LOHHLA/QP, the mean RSS was 2.253, and the median was 0.4148. On the one hand, MAPSCE finds solutions with a lower RSS, suggesting that they are a better fit for the experimental data. On the other hand, the difference between the mean and median RSS for mapping subclonal events between both tools showed that MAPSCE's mapping algorithm provides enhanced consistency and reliability. This is supported by the mean value being closer to the median, suggesting more robust and stable results. Certain HLA alleles determined by LOHHLA/QP as subclonally lost had a significantly higher residual sum of squares above 2, compared to the median of 0.04148. Manual inspection of these cases showed that MAPSCE considered them to lack sufficient statistical support, the null branch was not rejected and therefore no subclonal event was called.

Interestingly, alleles identified as clonal LOH by MAPSCE were always determined to have a LOH event by LOHHLA/QP, albeit in four cases LOHHLA considered the event subclonal. Furthermore, most disagreements between the tools arose when one tool classified an event as subclonal and the other determined it to have no LOH. This discrepancy occurred 10 times for MAPSCE and 9 times for LOHHLA/QP. These 9 cases classified as subclonal by LOHHLA were labelled as no LOH in MAPSCE due to the lack of statistical support. This included 3 cases where the best branch's BICs were all higher than the null's BIC and 6 cases where the Bayes Factors comparison showed that all results were equally good, leading to the null not being rejected.

Chapter 4 demonstrated that the tool performed best in patients with more than two regions sequenced. In this analysis, we ran MAPSCE with default settings (i.e. without any specific constraint, with bootstrapping for cases with more than two regions and no bootstrapping for two regions). We assessed whether disagreements between LOHHLA/QP and MAPSCE were enriched in cases with two regions only. Among all discrepancies, only two alleles deemed subclonally lost according to MAPSCE and unaffected according to LOHHLA/QP had two regions sequenced. All other cases corresponded to patients with more than two regions sequenced.

We divided the alleles into two groups based on the clonality determination using both tools: matching alleles, indicating the same clonality (e.g. subclonal LOH for both MAPSCE and LOHHLA), and mismatching alleles, indicating different clonality (e.g. subclonal LOH for MAPSCE and clonal LOH for LOHHLA. These groups were then tested for confounding factors using a t-test. While the matching alleles corresponded to samples with a higher purity, the difference was not significant (p-value of 0.060). There was also no significant difference in the number of clones (p-value of 0.920) or the number of regions (p-value of 0.121) between the two groups. There was no significant difference in the proportion of genome duplication categories between patients with matching alleles and those with mismatching alleles (Fisher's exact test; p-value of 0.317).

These results demonstrated that MAPSCE's automated detection of subclonal events exhibited comparable performance to the heuristic approach employed by LOHHLA/QP in the TRACERx 100 dataset. While it is possible to argue in favour of either methodology in cases of disagreement, it is noteworthy that both tools agreed on the vast majority of alleles in terms of clonality determination. Importantly, MAPSCE provided additional statistical support for the results and more detail for the user to interpret and understand the mapping. All things considered, MAPSCE's clonality determination proved to be a more robust approach due to its automatic nature, as well as its lower and closely aligned mean and median RSS values.

5.2.2 Biallelic inactivation of tumour suppressor genes

Tumour suppressor genes play the crucial role of guarding the genome against replication errors underlying the tumour's ability to proliferate (Hanahan and Weinberg 2000; 2011). The loss of function of tumour suppressor genes is pivotal in cancer initiation and underlies other fundamental hallmarks of cancer, including the evasion of apoptosis and the unlimited replicative potential of cancer (Hanahan and Weinberg 2000; 2011). Knudson (1971) demonstrated the two-hit theory, in which each copy of the TSG needed to be affected independently for initiation of retinoblastoma. Cavenee et al. (1985) further identified that mutations involving the *RB1* tumour suppressor gene on chromosome 13 were responsible for the development of retinoblastoma. The biallelic epigenetic inactivation of *RASSF1* TSG was also linked to the development of medulloblastoma (Lusher et al. 2002). While biallelic inactivation of tumour suppressor genes has been identified as a driving mechanism in multiple cancer types, there is still a lack of this analysis at the subclonal level (Hamano et al. 2002; Thanendrarajan et al. 2017).

We used MAPSCE's functionality to explore subclonal biallelic inactivation of tumour suppressor genes in the NSCLC TRACERx 100 dataset. In this analysis, only somatic mutations and copy number events were considered.

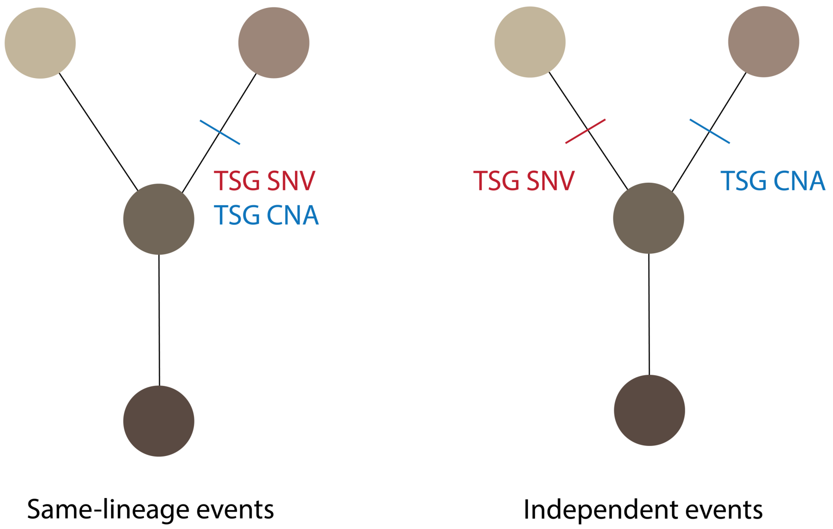


Figure 5-2 The difference between events on the same lineage and independent events.

We used the catalogue of driver genes from IntOGen (2020.02.01 release) to identify a list of 584 tumour suppressor genes (Martínez-Jiménez et al. 2020) (Table S5-1). All losses of TSGs, along with all mutations, were mapped on the tumour trees built using the SNVs. In total, we found 624 cases of biallelic inactivation of TSGs at the regional level. However, subclonal analysis revealed that 147 cases (23.6%) were independent events, where both events affected the same patient but in two different lineages (Figure 5-2). 340 cases were “double clonal”, which we defined as having both clonal SNV and clonal loss of a TSG. Since MAPSCE was designed to map subclonal events specifically, we focused on the events with subclonal loss of TSGs mapped using our tool. In total, there were 151 cases of biallelic inactivation of TSGs with a subclonal loss. We classified these 151 cases into same-lineage events (16 cases) (Table S5-2), where both events happened in the same lineage, thus leading to the true subclonal biallelic inactivation of a TSG, and independent events (135 cases) (Figure 5-2). The rare occurrence of same-lineage events including a subclonal loss of a TSG can be explained by the fact that the majority of the events affecting TSGs conferred a growth advantage and became clonal after a clonal sweep.

We tested if the same-lineage events were associated with any particular cancer subtype compared to the independent events using the histological classification of the patient. However, there was no relationship between same-lineage or independent events in both lung adenocarcinoma and lung squamous carcinoma (Fisher's exact test; p-value of 0.775). We used the MGSA R package to test for the gene set association for the LUAD and LUSC drivers. The independent events were more associated with the LUSC rather than LUAD drivers (0.1396 and 0.0790 posterior probabilities, respectively), although both posterior probabilities were still lower than 0.5. The same-lineage events showed no association (0 posterior probability) for either LUAD or LUSC drivers.

We reasoned that if the biallelic inactivation of TSGs provides a growth advantage, it could in turn lead to an acceleration in the evolution of the tumour and an increase in mutational burden. To test this hypothesis, we compared the nonsynonymous to synonymous mutation ratio of the lineage after subclonal biallelic inactivation of a TSG (affected lineage or clone) to its sister lineage sharing a common ancestor (sister lineage or clone) using the dNdS R package (Martincorena et al. 2017) (Figure 5-3). This comparison aimed to ascertain whether there was an evident increase in clones affected by the biallelic inactivation of a TSG compared to the unaffected sister clones.

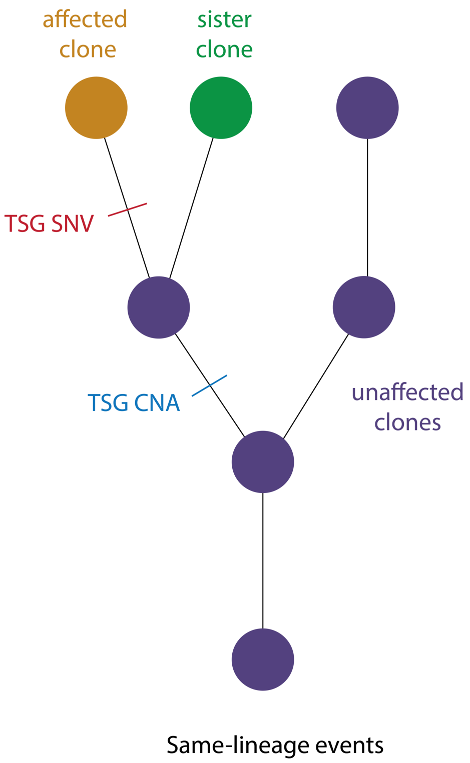


Figure 5-3 An example of a same-lineage event with defined: unaffected (purple), sister (green) and affected (orange) clones.

Clones affected by the same-lineage events generally had a lower dN/dS (1.079) than sister clones (1.161), however, there was no significant difference between the two sets (p-value of 0.682). That said, we observed five cases, where there was an increase of dN/dS in the affected lineage (mean dN/dS of 1.48 for these five cases) compared to their respective sister lineage (mean dN/dS of 0.98 for these five cases). The increased dN/dS of the affected clones in these cases suggest that these affected clones were likely to be in the process of outcompeting their sister clones.

Additionally, to examine the potential increase in the mutational burden following a biallelic inactivation of a TSG we compared the number of mutations per clone of the lineage after subclonal biallelic inactivation of a TSG (affected lineage or clone) to its sister lineage sharing a common ancestor (sister lineage or clone) (Figure 5-3). We hypothesised that starting from the same genomic origin, there would be an increase in the number of mutations after the biallelic inactivation of a TSG. This was determined by counting the ratio of the average number of mutations of each clone in the lineage and normalised against the whole tree. The clones in the affected lineage generally included a higher average number of mutations (0.125) than clones in their sister lineage (0.062). However, no significant difference was observed between the two datasets (p-value of 0.108).

The clones affected by the subclonal biallelic inactivation of a TSG exhibited generally lower dN/dS and a considerably higher number of mutations. However, no significant difference was found in the analysis. These findings suggest that a subclonal inactivation of a TSG leads to an increase in the mutational burden measured as the number of mutations per clone. However, the dN/dS analysis on the same-lineage events also showed that the subclonal biallelic inactivation of a TSG is not generally a cause for positive selection.

Given the low sample size of the patients with subclonal same-lineage biallelic inactivation of tumour suppressor genes, it was important to look into specific examples of how MAPSCE could be utilised to investigate a particular tumour's evolution. Thus, we analysed three examples of biallelic inactivation of the *BRCA2* gene in the cohort. *BRCA2* plays a crucial role in DNA damage repair (Cheng et al. 2016; Yoshida and Miki 2004), and the inactivation of the *BRCA2* and *BRCA1* genes has been strongly associated with the mutational signature SBS3 (Alexandrov et al. 2013). The biallelic inactivation of the *BRCA2* gene was mapped on the tumour evolutionary trees to compare the dN/dS, the number of mutations and ultimately, the rise of the mutational signature SBS3 between clones affected and unaffected by the biallelic inactivation of *BRCA2* (Figure 5-4).

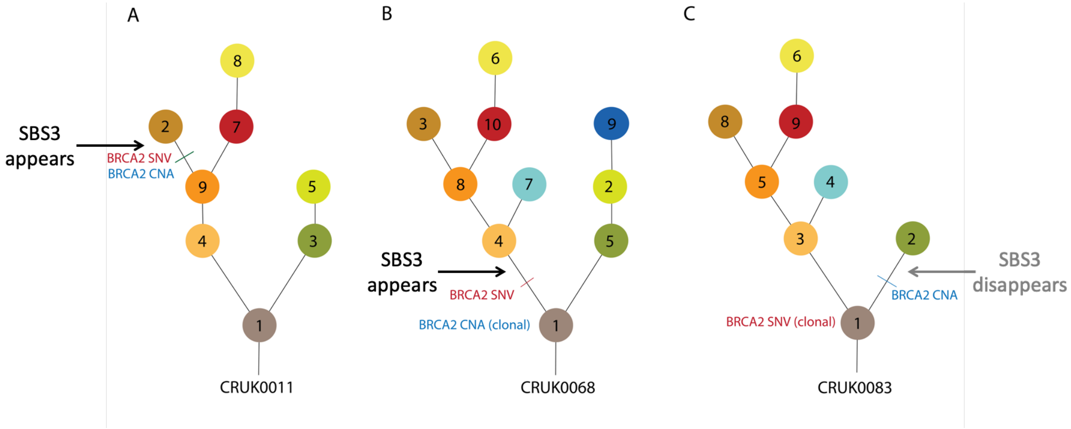


Figure 5-4 Copy number and mutational changes for the three patients with subclonal biallelic inactivation of BRCA2 of the NSCLC TRACERx 100 dataset. Panels A-C show different patient trees, A – CRUK0011, B – CRUK0068, and C – CRUK0083.

We compared the effects of the biallelic inactivation of *BRCA2* on the dN/dS in affected clones and their respective sister lineages. Patients CRUK0011 and CRUK0068 recorded a decreased dN/dS in the lineages affected by the biallelic inactivation of *BRCA2* compared to their sister lineages (0.90 vs 1.10, and 1.11 vs 1.15, respectively). Conversely, there was a large increase of dN/dS in affected (1.58) compared to sister clones (1.06) in patient CRUK0083. Looking at the tumour burden, all three patients showed an increased number of mutations in the affected clones compared to sister clones (CRUK0011: 0.155 vs 0.0249; CRUK068: 0.0628 vs 0.0520; and CRUK0083: 0.840 and 0.155 respectively).

We mapped the mutational signatures on the subclonal level using the deconstructSigs R package (Rosenthal et al. 2016). We performed this analysis by comparing the affected clones (Figure 5-3) to both all unaffected clones, i.e. sister and parental clones combined (Figure 5-3) as background. We expected the mutational signature SBS3 to not be present in the rest of the tree and only appear after the subclonal biallelic inactivation of *BRCA2*. In two out of three patients (CRUK0011 and CRUK0068), the mutational signature SBS3 could be specifically identified after the biallelic inactivation of the *BRCA2* gene (Figure 5-4A, 5-4B). Patient CRUK0068 had a clonal LOH of *BRCA2* followed by a subclonal mutation of the gene, while patient CRUK0011 had both the loss of *BRCA2* and mutation of the gene occurring on the same subclonal branch (Figure 5-4A, 5-4B). For patient CRUK0083, the mutational signature SBS3 was detected in the root of the tree already and disappeared after the second hit on the *BRCA2*, which was a loss (Figure 5-4C). In this patient, the mutation of *BRCA2* was a clonal event occurring on the root of the tree, while the loss of the gene was a subclonal event. These results suggest that the mutational signature SBS3 is more dependent on the mutation of the *BRCA2* gene rather than the CNA or the biallelic inactivation of the gene.

The dN/dS results did not show a clear example of positive selection for the biallelic inactivation of the *BRCA2* gene. However, the consistent increase in the number of mutations in affected clones compared to their sister lineages suggests that *BRCA2* potentially leads to an increased tumour burden (Zámborszky et al. 2017)**.** Nevertheless, these results illustrate the efficacy of MAPSCE and how it allows for an in-depth analysis of the different causes driving the evolution of a tumour.

5.2.3 Mutual exclusivity of antigen-presentation machinery

HLA LOH has been identified as an immune escape mechanism (McGranahan et al. 2017) which inhibits neoantigen presentation. During the development of MAPSCE's framework, a question was raised regarding the presence and timing of the subclonal homozygous loss of the other genes involved in antigen presentation (AP). Specifically, we were interested in whether these losses occurred on sister branches opposite to the subclonal loss of heterozygosity of HLA or within the same lineage as the HLA LOH event. A consistent pattern of either co-occurrence or mutual exclusivity of certain AP genes involved in antigen presentation machinery could help identify novel evolutionary constraints of NSCLC development. To test this hypothesis, MAPSCE was used to map the subclonal losses of antigen presentation genes from Gene

Ontology (GO:0019882) on the tumour trees of the NSCLC TRACERx 100 dataset.

In this section, unless explicitly stated otherwise, “MAPSCE” refers to MAPSCE (v.0.5.0).

MAPSCE identified a total of 12 genes, whose subclonal loss always occurred on a sister branch to the subclonal HLA LOH event (Figure 5-5). We examined these genes for their involvement in the antigen processing and presentation pathway based on the KEGG dataset (Figure 5-6). Among the identified genes, only three (*TAP1*, *TAP2* and *TAPBP*) were found to be associated with the transport of antigenic peptides across the endoplasmic reticulum, as previously reported (Maeurer et al. 1996). The consistent occurrence of these subclonal events on a sister branch is indicative of mutual exclusivity. Interestingly, these three genes all corresponded to the major histocompatibility complex (MHC) class I, akin to the examined HLA genes. This mutual exclusivity could be explained by further events affecting MHC class I not conferring additional growth advantage.

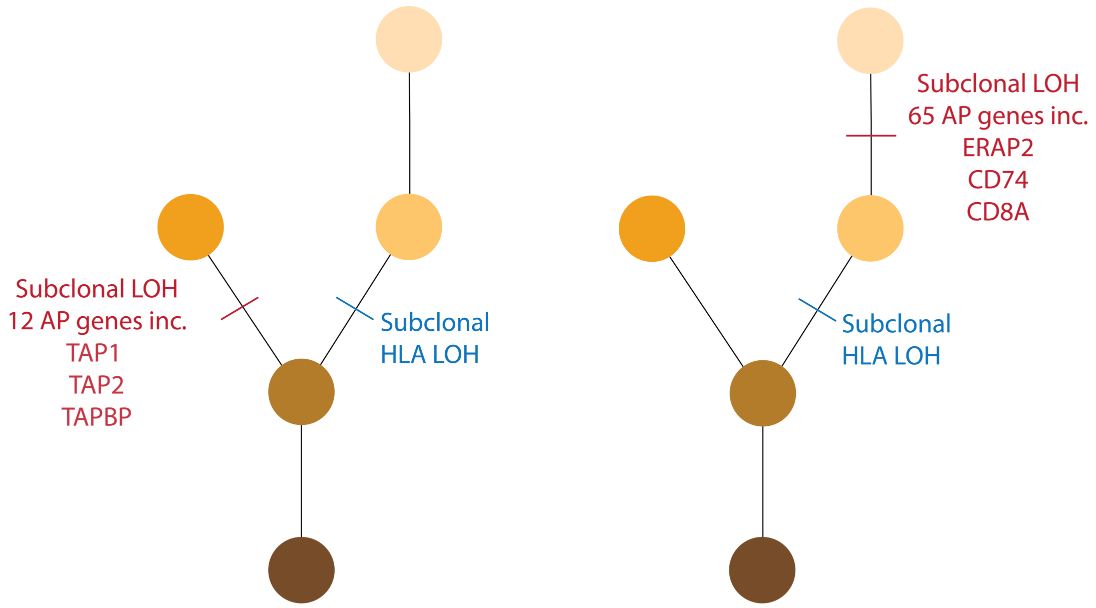


Figure 5-5 Subclonal LOH of antigen presentation genes on a sister branch (left) and in the same lineage (right) relative to subclonal HLA LOH.

MAPSCE also identified 65 other antigen presentation genes whose subclonal losses always occurred within the same lineage following a subclonal HLA LOH event (Figure 5-5). These genes included but were not limited to, *ERAP2*, *CD74*, *CD8A*, *PCNX* and *PSMA6*. Importantly, *CD8A* is normally active in the CD8+ T-cells as it plays a crucial role in facilitating antigen recognition and binding. The losses of *CD8A* in our results could have been a spurious result or passenger losses and were not relevant to the tumour cells. To explore their functional relevance, these genes were also examined for their involvement in the antigen processing and presentation pathway based on the KEGG dataset (Figure 5-6). Notably, all of these losses were consistently subclonal and occurred after an HLA LOH event. Interestingly, these co-occurring events affected genes primarily corresponding to the major histocompatibility complex (MHC) class II. This finding suggests that additional events affecting the MHC-II in addition to the HLA LOH (MHC-I) could be necessary for an improved immune escape mechanism.

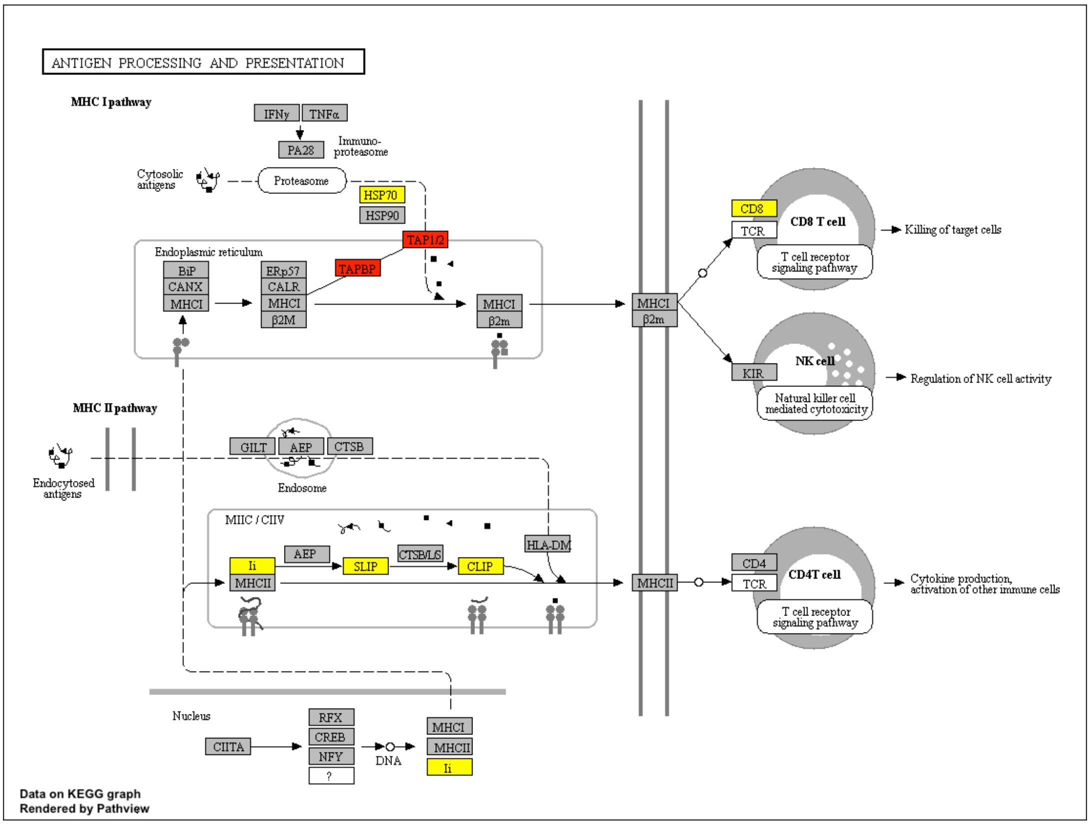


Figure 5-6 Genes with homozygous losses identified by MAPSCE and their association with the antigen processing and presentation pathway based on the KEGG dataset. Red labels show proteins coded by genes that were always lost on a sister branch to HLA LOH (mutually exclusive). Yellow labels show proteins coded by genes that were always lost within the same lineage to HLA LOH (co-occurring). The figure was made using R Package pathview.

In summary, MAPSCE facilitated the mapping of subclonal losses in antigen presentation genes to compare their timing with that of subclonal HLA LOH events. These results highlight the potential of mapping subclonal copy number events on the SNV-based tumour evolutionary trees to identify examples of parallel evolution and uncover novel selective forces shaping branched tumour evolution.

5.2.4 Allele-specific expression in NSCLC TRACERx

Transcriptomic variation is another major contributor to intratumour heterogeneity which influences tumour progression, therapy, and patient outcomes. Studies have shown the altered expression of specific cancer driver genes in metastatic melanoma (Tirosh et al. 2016; Rambow et al. 2018), glioblastoma (Neftel et al. 2019) and lung cancer (Biswas et al. 2019).

A recently published TRACERx study demonstrated a comprehensive multi-region analysis of the transcriptomic diversity in the 421 TRACERx NSCLC dataset (Martínez-Ruiz et al. 2023; Frankell et al. 2023). The study described the transcriptomic landscape of the TRACERx 421 cohort, distinguishing between the CN-dependent allele-specific expression (ASE) caused by genomic alterations, and CN-independent ASE linked to epigenomic variation. Specifically, patient CRUK0640 showcased the expression change of one of the two alleles of the *FAT1* gene in two different regions, where a CN event explained the downregulation in one region (CN-dependent ASE). However, no CN event was found in the other region, suggesting an epigenetic modification drove the downregulation of the allele in that case (CN-independent ASE). The phylogenetic analysis demonstrated the independent evolution of the two regions leading to the same phenotype of a loss of the *FAT1* gene through different mechanisms.

We compared the subclonal allele-specific gene expression changes with the subclonal genomic alterations using MAPSCE. This analysis aimed to identify both CN-independent and CN-dependent expression changes. The work presented in this section was performed with MAPSCE in its developmental stage (v0.5.0).

MAPSCE revealed a subclonal CN loss of the *FAT1* gene on branch 3 for allele A in patient CRUK0640 (Figure 5-7). Subsequently, we analysed the subclonal allele-specific expression changes based on changes in the RNA read counts. We observed allele-specific expression changes of allele A on branch 4 (best result, subclonal), branch 3 (second-best result, subclonal), branch 7 (third-best result, subclonal) and branch 1 (fourth-best result, clonal). Notably, the second-best result of MAPSCE's mapping of expression changes aligned with its mapping of copy number alterations, illustrating an example of CN-dependent ASE. Additionally, MAPSCE showed a subclonal ASE change on branch 4 which was not followed by a subclonal CNA, demonstrating an example of CN-independent ASE. Nominally, MAPSCE assumes that only one event has happened. However, in this case, there are possibly more than one event.

MAPSCE (v0.5.0) allowed for more than one good result without consolidating the result with consensus mapping. Thus, all four of the best results for mapping expression changes of the allele A were deemed sufficiently good at the time of analysis to be considered a potential solution.

These results are consistent with previously reported parallel evolution in patient CRUK0640, in which CN-dependent and CN-independent ASE were found on two different regions, with convergence upon the loss of different alleles of *FAT1* through genomic and transcriptomic means (Martínez-Ruiz et al. 2023).

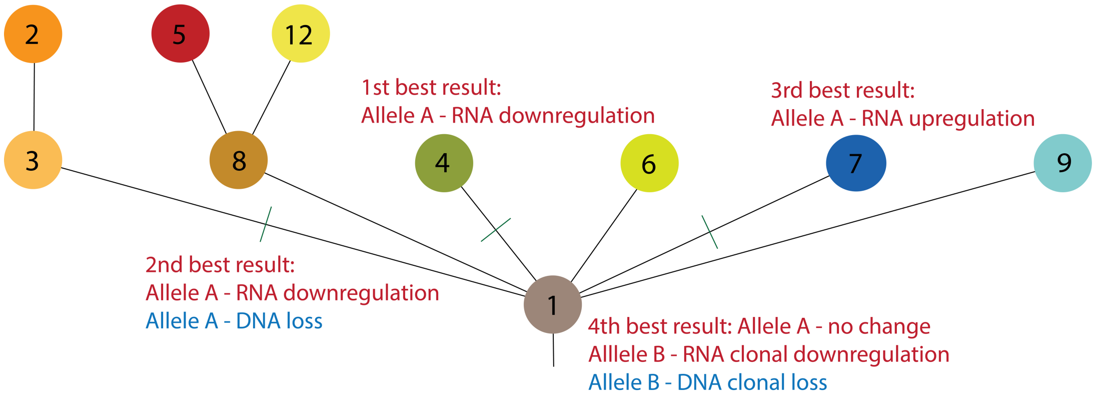

2
5
12
1st best result:
Allele A - RNA downregulation
3rd best result:
Allele A - RNA upregulation
3
8
4
6
7
9
2nd best result:
Allele A - RNA downregulation
Allele A - DNA loss
1
4th best result: Allele A - no change
Alllele B - RNA clonal downregulation
Allele B - DNA clonal loss

Figure 5-7 Mapped copy number changes (blue) and ASE changes (red) of the FAT1 gene on the tumour tree of the patient CRUK0640.

We analysed all pairs of alleles of 152 genes that exhibited CN-dependent ASE only, following the methodology outlined by Martínez-Ruiz et al. (2023).

This analysis aimed to determine if these genes displayed co-occurrence of CN and ASE changes on the same branch of a tumour tree. MAPSCE was utilised to map the CN events (using the DNA reads) and the ASE changes (using the RNA reads) separately for each gene. In addition to checking the co-occurrence of the subclonal events, the directionality of the events was interrogated. This involved examining whether a loss corresponded with downregulation and a gain with upregulation. We repeated this analysis using various approaches, including:

- mapping allele-specific RNA reads compared to CN changes (as in the aforementioned *FAT1* analysis),
- mapping purity-adjusted allele-specific RNA reads compared to CN changes,
- non-directional mapping the BAF of the purity-adjusted RNA reads compared to CN changes

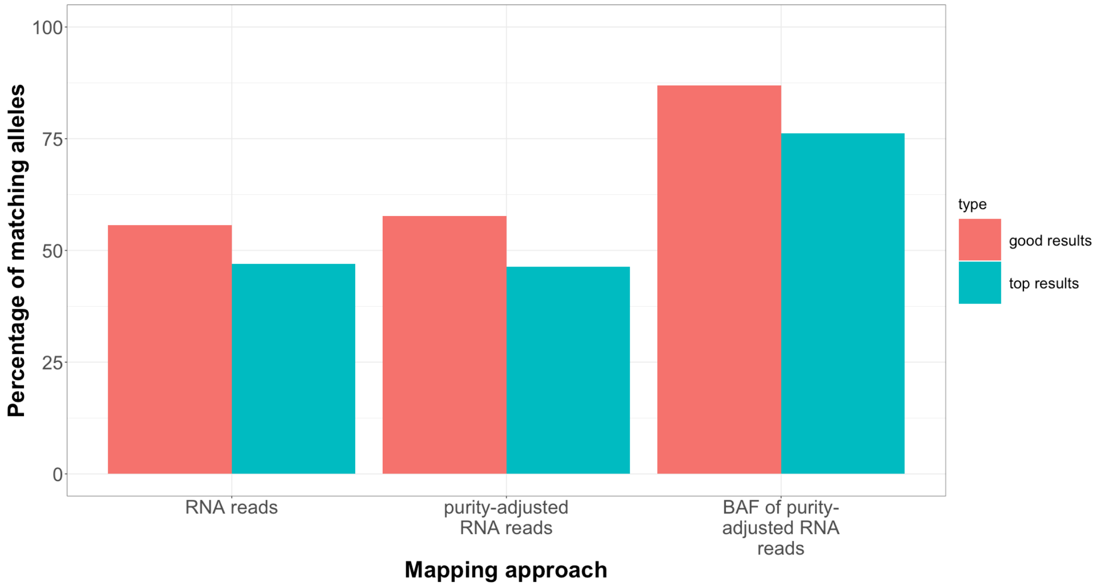


Figure 5-8 Percentage of matching alleles for different approaches to mapping CN changes and ASE changes in genes with CN-dependent ASE. The good results indicate all potential good results of MAPSCE, while the top results denote only the best branch selected for every mapping.

The results were split for all good results and only the best results (top results) of MAPSCE. For all good results of both CN and ASE mapping, an allele was considered a match between the CN and ASE change mappings if the alterations were mapped within one of the potential good results for either mapping. In most cases, the top result matched the CN change, but an additional 10% agreement could be found when using all good results (Figure 5-8). Looking into top results specifically aimed to assess the precision of MAPSCE's mapping at the time.

There was minimal difference between the percentage of matching CN and ASE changes between the purity-adjusted RNA reads and raw RNA reads. Correcting the expression data for sample purity had little effect when mapping ASE changes as RNA reads (55.7% for raw RNA reads to 57.7% for purity-adjusted RNA reads).

Lastly, when using non-directional BAF of the purity-adjusted reads instead of RNA reads, the transcriptomic and genomic data matched most of the time (86.9% matched with good results of MAPSCE considered and 76.2% with only the top results). MAPSCE consistently mapped both genomic and transcriptomic events on the same branch of the tumour tree for CN-dependent ASE.

The results of mapping the losses and ASE changes of *FAT1* suggest potential parallel evolution between the sister lineages of the patient CRUK0640 tree leading to the loss of *FAT1* through genomic and transcriptomic changes, consistent with the findings of Martínez-Ruiz et al. (2023). However, MAPSCE assumes there is at most one subclonal event for each gene for a particular tree. In this case, the different solutions of MAPSCE suggest the potential presence of more than one subclonal event. Additionally, this analysis demonstrated MAPSCE's ability to validate the CN-dependent ASE by mapping the genomic and transcriptomic changes separately on a tumour evolutionary tree. In summary, these results show how MAPSCE can be utilised to integrate genomic and transcriptomic data to identify examples of parallel evolution.

5.3 Conclusions

5.3.1 Summary of findings

This chapter highlighted the advantages of MAPSCE's integrative approach to multi-omics data. Using the tool in the analysis of real datasets showcased its ability to identify potential examples of parallel evolution and to provide validation for novel evolutionary principles.

Firstly, we revisited the previous analysis presented in McGranahan et al. (2017) and compared MAPSCE to LOHHLA in mapping subclonal HLA LOH events on the tumour trees of the NSCLC TRACERx 100 dataset. Since both tools employ quadratic programming for their mapping algorithm, there was a substantial level of agreement between the two methodologies. One of the main differences between both tools was that MAPSCE offers additional statistical support for its results, providing the user with more information to interpret the results. MAPSCE's automated clonality determination resulted in lower mean and median RSS values compared to results obtained from LOHHLA's heuristic approach. Thus, MAPSCE proved to be a more robust and consistent tool for clonality determination. Unlike LOHHLA's mapping approach, MAPSCE is not limited to mapping subclonal LOH events, but also allows for the integration of other types of events, showcasing its broader functionality.

The biallelic inactivation of tumour suppressor genes has been shown to drive tumour initiation for numerous cancer types, such as retinoblastoma (Knudson 1971; Cavenee et al. 1985), medulloblastoma (Lusher et al. 2002), prostate cancer (Cheng et al. 2016), sporadic renal cell carcinoma (Hamano et al. 2002), multiple myeloma (Thanendrarajan et al. 2017) and many others. However, genes that on the regional level appear to be biallelically inactivated can have both events on different lineages, suggesting potential parallel evolution affecting this particular TSG. Expanding on the findings of the previous analysis, we utilised MAPSCE to map the subclonal events affecting the tumour suppressor genes in the NSCLC TRACERx 100 dataset. In the NSCLC TRACERx 100 dataset, the majority of the events affecting TSGs were clonal rather than subclonal. Unfortunately, the scarcity of subclonal events affecting tumour suppressor genes means that without an even larger cohort, there is not enough statistical power to identify broader evolutionary patterns. Thus, we focused on three examples of subclonal biallelic inactivation of *BRCA2*. The slight increase in the number of mutations in the affected clones could be indicative of positive selection for the double hit of *BRCA2*. However, there was no consistent and significant evolutionary pattern identified, except from an increased tumour burden following biallelic inactivation of *BRCA2*. Lastly, we focused on the mutational signature SBS3, which is strongly associated with the inactivation of *BRCA1* and *BRCA2* (Alexandrov et al. 2013). In these patients, the emergence of signature SBS3 was related to mutations on *BRCA2* rather than to the loss of the gene. The clonal status of the majority of events affecting TSGs suggests that a single subclonal event affecting tumour regions confers a sufficient growth advantage. This is consistent with the hypothesised parallel evolution leading to the loss of HLA and *FAT1* in NSCLC (McGranahan et al. 2017; Martínez-Ruiz et al. 2023).

To understand the evolutionary processes underlying antigen presentation machinery, we compared the timing of subclonal losses of AP genes to the timing of subclonal HLA LOH events. Using MAPSCE, we identified a total of 12 genes that always occurred on a sister branch to the HLA LOH event, indicating mutual exclusivity. These genes, akin to the HLA genes examined, affected the MHC class I. The mutual exclusivity of these events suggests that further losses of the MHC class I do not confer additional growth advantage. Moreover, we discovered 65 other AP genes that always appeared within the same lineage following a subclonal HLA LOH, demonstrating a pattern of co-occurrence. These co-occurring losses affected genes corresponding to the MHC class II. These findings suggest that the HLA LOH (MHC-I) requires additional events hampering MHC-II to provide an improved immune escape mechanism.

Lastly, we used MAPSCE to track the transcriptomic and genomic changes along the tumour evolutionary tree of a single patient in the NSCLC TRACERx 421 cohort. The analysis of allele-specific expression changes in patient CRUK0640 revealed potential evidence for parallel evolution, with independent lineages developing both CN-dependent and CN-independent ASE of the *FAT1* gene on sister branches. This result is consistent with the findings of other studies (Martínez-Ruiz et al. 2023). Additionally, in our attempt to match the transcriptomic and the genomic changes for genes with CN-dependent ASE using MAPSCE, we observed varying degrees of success. These challenges can be attributed to the lack of consensus mapping in this version of MAPSCE (v0.5.0). Nevertheless, these findings showcase MAPSCE's novel ability to integrate the subclonal genomic and the transcriptomic data within the context of tumour evolutionary trees, which, to the best of my knowledge, has not been previously attempted systematically.

5.3.2 Limitations and future work

The analysis of the antigen presentation machinery in the NSCLC TRACERx 100 dataset and the mapping of subclonal allele-specific expression and copy number changes in a patient of the NSCLC TRACERx 421 cohort were performed using an early version of the tool (v0.5.0). While the bootstrapping of BICs, the conversion to Bayes Factors, and the comparison to the null hypothesis were already present, there was no consensus mapping to integrate the results in agreement in this version (v0.5.0).

Furthermore, MAPSCE v0.5.0 adopted a more lenient approach and considered a broader range of results as potentially valid. The Bayes Factors comparison involved sequentially comparing the top result's BF to the second-best result's BF, the second-best's to the third-best's, and subsequent pairs, to evaluate the relative strength of each result. This led to less precise mapping as more results were deemed potentially viable compared to the latest version of MAPSCE (v1.0.0), where all of the branches' Bayes Factors are individually compared to the top branch's BF.

At the time, no confounding factors were identified, and there was no separate mode for the patients with two regions sequenced. The quadratic programming in MAPSCE still constrained both the copy number states before and after to non-negative.

In future work, it would be valuable to repeat these analyses using the latest version of MAPSCE (v1.0.0). This updated version includes consensus mapping, which allows for enhanced mapping precision, addressing some of the limitations observed in the previous iteration of the tool (v0.5.0). The improved mapping precision in the version of the tool released on GitHub (v1.0.0) would be beneficial for accurately determining the sequential order of events and providing stronger evidence of parallel evolution, mutual exclusivity and co-occurrence.

The mutational signature analysis on the three example cases of *BRCA2* biallelic inactivation suggests that the mutational signature was more dependent on the mutation rather than the CNA or the biallelic inactivation event. However, specifically in the case of patient CRUK0083 we recorded a disappearing mutational signature SBS3. This analysis was performed by determining the relative contribution of each mutational signature. In patient CRUK0083, the mutational signature SBS3 decreased from a relative contribution of 7.3% to 0%, while mutational signatures SBS6 and SBS7 increased considerably from 8.7% and 0% to 22% and 28.8% respectively. Improving this analysis requires the quantification of the absolute contribution of the mutational signature SBS3 in each patient. In the future, tools for tracking of mutational signatures along the tree could also be utilised for validation of the results (Miura et al. 2022).

The mapping of subclonal losses of antigen presentation genes with respect to subclonal HLA LOH provided only preliminary insights into potential mutual exclusivity between losses of *TAP1*, *TAP2*, *TAPBP* and HLA LOH, and highlighted that the additional losses of genes affecting MHC-I did not confer added growth advantage. Conversely, the co-occurrence of losses of genes corresponding to MHC-II and HLA LOH events (MHC-I) suggests that the immune response acts as a key selective pressure driving mutagenesis to hamper the MHC-II pathway as well. However, to establish definitive proof of parallel evolution or co-occurrence, future analyses could include a larger sampling size and incorporate more robust statistical evidence. Furthermore, there are no tools available to examine the mutual exclusivity and co-occurrence at the subclonal level to formally test this hypothesis.

When examining ASE changes and how they co-occurred with the genomic changes on the tumour trees, it is important to consider the differences between the definition of a CN-dependent and CN-independent ASE according to the methodology of Martínez-Ruiz et al. (2023) compared to the mapping results obtained by MAPSCE on the tumour trees. We integrated the regional data to track the subclonal changes of expression along the patient's tumour tree. In contrast, Martínez-Ruiz et al. (2023) considered each sampled region independently, analysing CN losses and ASE changes within each region separately. As a result, MAPSCE did not detect any copy number events for certain genes, that were classified as exhibiting CN-dependent ASE by Martínez-Ruiz et al. (2023). Furthermore, MAPSCE works under the assumption that there is only one event occurring per gene. Using MAPSCE, it is possible to extend this analysis to more complex scenarios.

The novelty of MAPSCE in identifying evolutionary processes on the subclonal level makes validation of the results challenging. Phylogenetic reconstruction using CNAs could provide a measure of comparison of the mapping results of MAPSCE in cases where the SNV-based and CNA-based trees would be in agreement. Cell culture experiments provide a potential experimental validation considering the controlled environment for modelling of the tumours. However, it is challenging to replicate the complex tumour microenvironment using a simplified cell culture system to model the evolutionary processes in cancer. In vivo lineage tracing experiments also offer the ability to study the developmental history and fate of individual cancer cells. Conversely, lineage tracing is still prone to sampling bias depending on the specific selection of cells within a heterogeneous tumour mass. The procedures involved in lineage tracing experiments can additionally disrupt the tumour microenvironment.

Lastly, the recently published NSCLC TRACERx 421 dataset (Frankell et al. 2023) presents an opportunity for MAPSCE's mapping algorithm to gain increased statistical power due to a larger sample size compared to the original 100 patients (Jamal-Hanjani et al. 2017). The continuous generation of new multi-sample multi-omics data provides abundant opportunities to showcase the tool's broad functionality, particularly in mapping other data types such as the transcriptomic and methylomic changes.

Chapter 6 Discussion

The work presented in this thesis focuses on the development, testing, and validation of MAPSCE (MAPping SubClonal Events), a tool designed for tracking subclonal events on tumour evolutionary trees. The aim of the project was to develop a novel computational tool, which would allow for the integration of the multi-region, multi-omics and onco-genomics data on the subclonal level in the context of tumour evolutionary trees. This chapter will summarise the work presented in this thesis, highlight the novel features and functionality of the tool, outline the current limitations of the methodology and the findings, as well as discuss the potential improvements in future work.

6.1 Summary and novelty of the findings

The development of the tool involved exploring and testing various methods that could be utilised for mapping subclonal events of tumour evolutionary trees.

Firstly, we reviewed the mapping algorithm for subclonal HLA LOH events from LOHHLA (McGranahan et al. 2017) to identify the limitations of the existing methodology. This approach employed heuristic clonality determination with quadratic programming to integrate CN and SNV data in the context of tumour evolutionary trees. Although novel at the time, LOHHLA's mapping of subclonal copy number events on SNV-based tumour trees lacked a measure of goodness of fit and was limited to subclonal copy number events following prior determination of their presence. The mapping of subclonal HLA LOH events on SNV-based tumour trees was later explored in high-grade serous ovarian cancer (Zhang et al. 2018). These proof-of-concept studies demonstrated the potential of integrating different types of events, the CN and the SNV data, in the context of a tumour evolutionary tree.

Chapter 3 further explored additional methods to expand mapping capabilities to other types of subclonal copy number events, like copy number gains, while also introducing a measure of goodness of fit for the results. Initially, non-negative least squares were considered as a potential replacement for quadratic programming in the branch test. However, we found that nnls could be presented as a form of quadratic programming minimisation (Equation 4) (Bro and Jong 1997), which meant it faced similar limitations as quadratic programming.

We explored a method to directly derive CCF from copy number data, to validate the results obtained from mapping subclonal copy number events. While this approach worked well for straightforward cases in its simplified form, its implementation became circular when applied to more complex events. The circularity stemmed from deriving CCF from copy numbers that were inferred using quadratic programming, which relied on the CCF in the first place. Further work is required to address the implementation of this approach.

Instead of finding an alternative to the quadratic programming in the branch test, we explored adding statistical support to its results. We first tested simulations of noise in artificial datasets to provide an objective measure of the method's performance. Initially, we attempted to implement Approximate Bayesian Computation to provide the posterior distributions for different branches as a measurement of the goodness of fit of the results. However, due to the computational intensity of ABC, we tested a pseudo-ABC with maximum likelihood estimation instead. Despite its accuracy, this approach required a high number of simulations for result stability. Even with a simpler sampling approach, implementing ABC or pseudo-ABC with MLE would require considerable computational resources for mapping CN of every gene of each patient in a larger cohort.

As an alternative to ABC, we explored bootstrapping with filtering of results. This approach allowed for model selection and measurement of the goodness of fit using RSS converted into Bayesian Information Criterion and Bayes Factors. Combining these features and consensus mapping to integrate the results in agreement improved the mapping precision of the tool. Furthermore, adjusting quadratic programming in the branch test allowed mapping other copy number events, including copy number gains. Together, these results helped shape the current framework of MAPSCE.

To evaluate the tool’s performance, we simulated various copy number events (Chapter 4), including amplification (two to three copies), duplication (two to four copies), homozygous loss (two to zero copies), heterozygous loss (two to one copy) and loss of heterozygosity (one to zero copies) events. Since there is no established gold standard method for mapping subclonal copy number events, we benchmarked MAPSCE against LOHHLA’s mapping approach (McGranahan et al. 2017) on the datasets with simulated copy number events. Since LOHHLA’s mapping approach was specifically designed to only map loss of heterozygosity events, the comparison between the two tools was limited to that particular type of event. The results showed that MAPSCE maintained a higher mapping accuracy for subclonal events than LOHHLA/QP, regardless of the number of regions. Notably, our tool consistently outperformed its competitor when mapping events in smaller cluster sizes, regardless of whether we considered the size of the simulated or the mapped cluster.

We demonstrated the novelty of the tool specifically in mapping other types of CN events, such as copy number losses extending beyond LOH. These included homozygous and heterozygous losses, copy number gains, duplications and amplifications. Across different cluster sizes, MAPSCE consistently maintained high mapping accuracy for subclonal events for events with more than two regions sequenced and with constraints for events with two regions sequenced.

Mapping copy number gains posed a challenge in determining the copy number of the most recent common ancestor (MRCA) or the root of the tree. However, providing the tool with the appropriate constraints to define the expected type of event it would encounter in the dataset considerably improved the mapping accuracy. What sets MAPSCE apart is its ability to integrate various copy number events with the SNV data in the context of a tumour evolutionary tree, an approach which has not been explored previously. Existing tools reconstruct phylogenies using the CN or the SNV data independently and compare the two individual trees for potential correlations (Malikic et al. 2015; Miller et al. 2014). Other tools integrate both SNV and CNA data in their phylogenetic reconstruction, however those tools re-analyse the data to provide their own interpretation. This approach is more computationally intensive and produces less versatile data (Prandi et al. 2014; Deshwar et al. 2015). MAPSCE integrates the output of state-of-the-art tools for deciphering the intratumour heterogeneity (Van Loo et al. 2010), making it compatible with other approaches.

Additionally, we validated the tool on real datasets, highlighting the novelty of its features and demonstrating its potential in addressing various biological questions. We compared MAPSCE and LOHHLA's mapping approach on the NSCLC TRACERx 100 HLA LOH dataset (Jamal-Hanjani et al. 2017; McGranahan et al. 2017). This comparison aimed to assess the accuracy of clonality determination of both tools. The results showed a high agreement of 90.3% between both methods when classifying HLA alleles as having either no LOH, clonal LOH or subclonal LOH. While both tools utilise quadratic programming to map subclonal events, LOHHLA adopts a heuristic approach for its clonality determination, whereas MAPSCE relies on an automated method based on its mapping results. The high agreement between both tools could be attributed to their shared reliance on the observed copy number either directly in the case of LOHHLA/QP or indirectly in MAPSCE's mapping algorithm. These results demonstrate that MAPSCE's clonality determination is, at the very least, comparable to that of another previously published study (McGranahan et al. 2017).

Furthermore, we compared the timing of the genes involved in antigen presentation machinery to the timing of subclonal HLA LOH events in the NSCLC TRACERx 100 dataset (Jamal-Hanjani et al. 2017). We identified 65 genes that, if lost, consistently appeared within the same lineage after a subclonal HLA LOH event. Those genes included *ERAP2*, *CD74*, *CD8A*, *PCNX* and *PSMA6*. The co-occurrence of the homozygous losses of these genes suggests that the immune system could be acting as a key selective pressure, driving further mutagenesis in antigen presentation machinery. The subclonal HLA LOH event could be insufficient in releasing that pressure, which led to additional events affecting the MHC class II. We also found a set of 12 genes whose losses consistently occurred on a sister branch to the subclonal HLA LOH event. The timing of these losses could suggest a potential mutual exclusivity between the events affecting the two lineages, ultimately leading to a loss affecting the APM. These 12 mutually exclusive genes corresponded to the MHC class I, akin to the examined HLA genes. However, to establish a pattern of mutual exclusivity or parallel evolution among those genes, additional statistical analysis, and a larger sample size are necessary. Taken together, in this analysis, we demonstrated MAPSCE's potential in uncovering novel evolutionary processes on a subclonal level.

Knudson's (1971) two-hit theory highlighted the crucial role of biallelic inactivation of the *RB1* tumour suppressor gene in tumour initiation and progression of retinoblastoma. This theory was further demonstrated in other cancer types, such as medulloblastoma (Lusher et al. 2002), sporadic renal cell carcinoma (Hamano et al. 2002), prostate cancer (Cheng et al. 2016), and multiple myeloma (Thanendrarajan et al. 2017). Thus, we identified the biallelic inactivation of TSGs on a subclonal level in the NSCLC TRACERx 100 dataset (Jamal-Hanjani et al. 2017). The biallelically inactivated TSGs were further categorized into those affected by the same-lineage events on a subclonal level and those with both events occurring on sister branches. We analysed these gene sets for correlations with different lung cancer types. Additionally, we examined the dN/dS, the number of mutations, and mutational signatures, specifically focusing on three cases of biallelic inactivation of *BRCA2*. Unfortunately, due to the limited sample size of subclonal biallelic inactivation of TSGs, we were unable to identify significant evolutionary processes within the dataset except from an increase in tumour burden defined by an increased number of mutations following a biallelic inactivation of *BRCA2*. The majority of events affecting TSGs were clonal, after a clonal sweep.

Finally, we compared the allele-specific expression and copy number changes on a tumour evolutionary tree for a single patient of the NSCLC TRACERx 421 cohort (Frankell et al. 2023). Consistent with a recently published study, MAPSCE identified evidence of potential CN-dependent and CN-independent ASE of *FAT1* for the CRUK0640 patient (Martínez-Ruiz et al. 2023). Furthermore, we compared the timing of the subclonal transcriptomic changes to the subclonal genomic events for the genes previously described as exhibiting CN-dependent ASE. The percentage of matching genomic and transcriptomic events varied depending on whether the tool mapped the RNA reads or the BAF of the RNA reads. Furthermore, there were notable differences in how CN events were defined in Martínez-Ruiz et al. (2023) compared to MAPSCE. The former identified subclonal CN events within each region independently, while MAPSCE utilised regional data to map the events and subsequently determine the clonality. It is important to note that this analysis was performed with the tool still in development (v0.5.0). We addressed the tool's issues with mapping precision later in the project by adding mapping constraints and consensus mapping to integrate results in agreement. Taken together, these results demonstrated MAPSCE's capability to integrate multi-region, multi-omics data in the context of a tumour evolutionary tree. To my knowledge, mapping of subclonal expression changes against copy number events on an SNV-based tree has not been attempted before, highlighting the novelty of MAPSCE. This further demonstrates how this approach can be extended to integrate more complex multi-omics data.

6.2 Limitations and future work

Throughout this thesis, each chapter has outlined specific limitations of the described work and proposed future approaches to improve the results, as indicated in the conclusions sections. This section will summarise those specific limitations, outline the general shortcomings of this work, as well as discuss the future directions opened up by the research presented in this thesis.

6.2.1 Single-cell sequencing and data availability

Single-cell sequencing has allowed for studying intratumour heterogeneity at the level of individual cells. The focus of cancer research has started shifting to integrate the scRNA-seq data (Navin et al. 2011; Tirosh et al. 2016; Wu et al. 2021; Ren et al. 2022; Schmiel, Thomas, and George 2022) with some studies integrating both single-cell RNA-seq and bulk RNA-seq data (Zhang et al. 2021). MAPSCE is specifically designed to track and integrate the multi-omics data on a subclonal level to decipher the intratumour heterogeneity in bulk sequencing data. The tool relies on multi-sample bulk sequencing data and has been thoroughly tested only for that use. Extending an algorithm beyond bulk sequencing data to include single-cell sequencing data has been done before with other tools, such as ASCAT (Van Loo et al. 2010). Implementing this feature would greatly improve MAPSCE's versatility.

One major limitation of the research included in this thesis is the data availability. While bulk sequencing data is prevalent in cancer research (Kuksin et al. 2021), there is still a lack of large multi-region cancer datasets, in which MAPSCE could be fully utilised for studying tumour evolution (Gerlinger et al. 2012; 2014; Zhang et al. 2014; Jamal-Hanjani et al. 2017; Frankell et al. 2023).

6.2.2 Validation of the tree space

During the phylogenetic reconstruction in the NSCLC TRACERx 100 dataset, we observed that multiple trees could apply to a single patient's data. The cluster CCFs used for reconstructing phylogenies were not consistently reliable, occasionally exceeding 100% for a particular lineage. In certain cases, tree inference can lead to several solutions. Throughout this project, we tested various versions of MAPSCE on these alternative trees, aiming to validate them based on their fit to the data. However, the phylogeny of the tree was not relevant as the key input for the tool was the CCF of the affected branch. Thus, the tool was not able to differentiate between different phylogenies. Since the integration of multi-omics data on the subclonal level relies on tumour evolutionary trees, validating the tree space would significantly improve our understanding of the tumour evolutionary dynamics. By addressing these limitations, we can further improve the accuracy of phylogenetic reconstruction, ultimately leading to enhanced mapping precision.

6.2.3 Exploration of noise in a real dataset

Chapter 4 of this thesis described the simulations of subclonal copy number events and the accompanying noise. Previous studies have explored various methods of simulating noise in the sequencing data or measuring the expected noise during subclonal deconvolution (Sloutsky et al. 2013; Barrett et al. 2017; Saunders et al. 2012). MAPSCE's acceptable thresholds of noise were determined to be up to 15 - 20% noise using the simulated datasets. To accelerate the testing of future tools developed for the analysis of sequencing data, further exploration of noise in real data is necessary. Understanding the average noise expected in the real data could further validate MAPSCE's results on the simulated datasets, determining whether the tool's resistance to noise is sufficient.

6.2.4 Improvements to the methodology

During the analysis of MAPSCE's performance on simulated subclonal copy number gains and losses in Chapter 4, the tool maintained higher mapping accuracy when given a constraint to determine the expected type of copy number events in the dataset. Providing the tool with the ability to automatically detect the types of copy number events in the dataset based on the observed copy number, and automatically setting an appropriate constraint would vastly improve the tool's performance and make it more robust for mapping different types of subclonal CNA. Currently, manual screening for the expected type of CNA in the dataset is required to maximize the tool's performance.

In Chapter 5, the mapping of allele-specific expression changes demonstrated the immense potential of MAPSCE in integrating different types of multi-omics data onto a single tumour evolutionary tree. However, to ensure consistent results, additional testing on simulated data and validation in real datasets is crucial for reliable mapping of expression changes on the tumour trees. To our knowledge, integration of methylation changes on SNV-based tumour evolutionary trees has not been attempted before. By extending the tool's functionality to track both subclonal expression and methylation changes together with CNA on the SNV-based trees, MAPSCE would become highly versatile in integrating multi-omics data in the context of tumour evolutionary trees.

6.3 Conclusion

Integration of multi-omics, onco-genomics data on a subclonal level within the context of tumour evolutionary trees allows for deciphering the evolutionary processes driving tumour progression. The lack of tools for tracking subclonal changes has hindered the ability to backtrack a tumour's evolutionary history. The research presented in this thesis has outlined the advantages and limitations of MAPSCE, a novel integrative tool for the precise mapping of subclonal events. By combining the genomic and transcriptomic data, this work provides the opportunity to study the diverse molecular alterations within different subclones. Extending this integrative approach to other multi-region, multi-omics and onco-genomics data holds immense potential for uncovering the evolutionary trajectories underlying intratumour heterogeneity across all cancer types.

Bibliography

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

Appendix

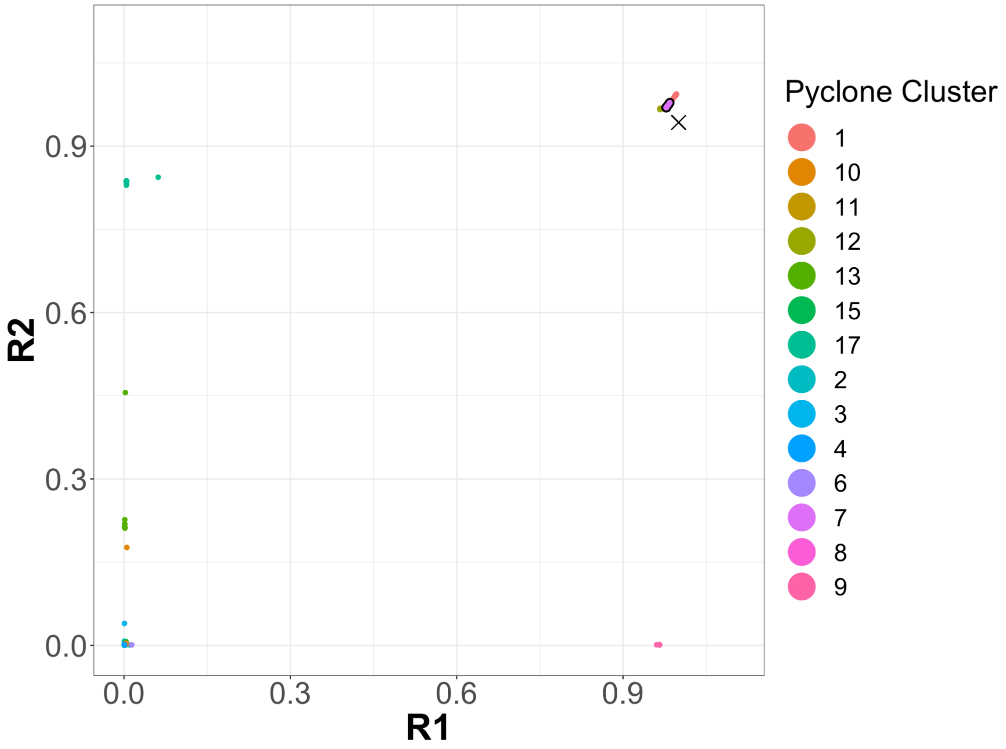


Figure S3-1. Example scatterplot of the mutations clustered based on their CCFs in region 1 (R1) and region 2 (R2). The different colours indicate the assigned cluster of the mutations during subclonal deconvolution using PyClone (defined as Pyclone Cluster in the legend). The cluster with a thick outline denotes the HLA LOH cluster inferred using MAPSCE. The cross shows the derived CCF_{CN}. Proximity of the cross to the HLA LOH cluster suggests matching results between MAPSCE and the derived CCF_{CN}.

Table S5-1. List of all tumour suppressor genes identified.

A1CF	ATP8A2	CBL	COL12A1	ELF4	FGFR4
ABCA6	ATR	CCDC6	COL1A1	ELL	FH
ABCF1	ATRX	CCDC85A	COL3A1	ELN	FHIT
ABI1	AXIN1	CCND1	COL6A3	EML4	FLCN
ABL1	AXIN2	CCND2	CR1	EP300	FLNA
ABL2	B2M	CCND3	CR2	EPHA2	FLT4
ACVR1B	BAP1	CCR7	CREBBP	EPHA3	FMR1
ACVR2A	BAZ1A	CD58	CSF3R	EPHA7	FN1
ADAMTS3	BCL10	CD70	CSMD3	ERBB2	FOXA1
ADAMTS8	BCL11B	CD79B	CTCF	ERBB3	FOXA2

ADCY8	BCL6	CDH1	CTDNEP1	ERBB4	FOXP1
AFF1	BCL9	CDH10	CTNNA2	ERCC3	FREM2
AFF3	BCL9L	CDH11	CTNND1	ERF	FUBP1
AHR	BCLAF1	CDH17	CTNND2	ERICH4	GATA2
AJUBA	BCOR	CDH18	CUL3	ESR1	GATA3
AKAP9	BCORL1	CDH9	CUX1	ESRRA	GCSAM
ALB	BCR	CDK12	CXCR4	ETV6	GFRA1
ALK	BIRC6	CDKN1A	CYLD	EXT2	GLI1
AMER1	BMP2K	CDKN1B	DAXX	EYS	GNA11
ANK1	BMP5	CDKN2A	DAZAP1	EZH2	GNA13
ANK2	BMPR1A	CDKN2C	DCC	FAM104B	GNAS
ANKRD36	BMPR2	CDX2	DDX3X	FAM135B	GRIA1
ANKRD36C	BORCS8-MEF2B	CEBPA	DGCR8	FAM46C	GRM3
APC	BRAF	CEP170	DICER1	FAM47C	GTF2I
APOB	BRCA1	CFHR5	DNAH9	FAM86B2	HELZ
AR	BRCA2	CHD2	DNMT3A	FANCA	HHLA3
ARHGAP35	BRD7	CHD4	DOCK3	FAS	HIP1
ARHGAP5	BTG1	CHEK2	DOT1L	FAT1	HIST1H1B
ARHGEF10	BTG2	CHRDL1	DPP7	FAT2	HIST1H1E
ARHGEF10L	BTK	CHRM4	DROSHA	FAT3	HIST1H2BL
ARHGEF12	C7orf55-LUC7L2	CIC	DSCAM	FAT4	HIST1H3B
ARID1A	CACNA1A	CIITA	DSG1	FBLN2	HLA-A
ARID1B	CACNA1D	CLIP1	DST	FBN2	HLA-B
ARID2	CAMTA1	CLTC	DTX1	FBXO11	HMCN1
ASXL1	CAPN5	CLTCL1	DUSP16	FBXW7	HNF1A
ASXL2	CARD11	CMTR2	DUSP2	FER1L6	HOXA11
ATF7IP	CARS	CNBD1	EBF1	FEZF1	HOXC13
ATG2A	CASP8	CNTN1	EHD2	FGF22	HOXD13
ATG7	CASZ1	CNTN5	EIF1AX	FGFR1	HSP90AA1
ATM	CBFB	CNTNAP2	ELF3	FGFR3	HSP90AB1
HTRA2	LTB	OR4C3	PTEN	SIN3A	TRAF2
HVCN1	LUC7L2	OR5L1	PTPN13	SIRPA	TRAF3
HYDIN	LYST	OR8H2	PTPN14	SLC34A2	TRIM33
ID3	LZTR1	OTOP1	PTPN6	SLC7A5	TRIM51
IDH2	MAML2	P2RY8	PTPRB	SLFN13	TRIO
IFNA6	MAN1A1	PABPC1	PTPRC	SLIT2	TRIP11
IFNGR1	MAP2	PAQR9	PTPRD	SMAD2	TRPV3
IGLL5	MAP2K4	PAX3	QKI	SMAD3	TSC1
IGSF21	MAP2K7	PAX5	RAD21	SMAD4	TSC2

IKZF1	MAP3K1	PBRM1	RANBP2	SMARCA1	TTLL2
IL1RL1	MAPK1	PCBP1	RAP1GAP2	SMARCA4	TTN
IL7R	MARK2	PCDH17	RAP1GDS1	SMARCB1	UBE2A
ING1	MAX	PCDH18	RARG	SMARCD1	UBE2D2
INO80	MB21D2	PCDH7	RASA1	SMO	UBR5
INPPL1	MECOM	PCDHB7	RASA2	SOCS1	UGT2B17
INSC	MED1	PCLO	RB1	SOX21	UNC80
IRAK1	MED12	PCMTD1	RBFOX1	SOX9	USH2A
IRF1	MEF2B	PDGFRA	RBM10	SP140	USP44
IRF8	MEN1	PDGFRB	RBM15	SPEF2	USP6
IRS4	MET	PDS5B	RBM38	SPEN	USP9X
ITGAE	MGA	PEG3	RCAN2	SPHKAP	VAV1
ITGAV	MSI2	PFAS	RECQL4	SPOP	VHL
ITGB6	MSN	PHF6	REG1A	SPTA1	WDR45
JAK1	MUC16	PIK3CA	RELA	SRRM2	WNK2
JAK2	MUC4	PIK3CB	RET	STAB2	WNK4
KANSL1	MYC	PIK3R1	RFX7	STAG2	WRN
KAT6A	MYH11	PIM1	RGPD3	STK11	WT1
KAT6B	MYH9	PITPNM2	RGS7	STX2	XRRA1
KDM3B	MYO5A	PLEKHG4B	RHOA	SUZ12	YLPM1
KDM5C	NBEA	PLXNB2	RHPN2	SYNE1	ZAN
KDM6A	NCOA1	PMS2	RIPK1	SYNE2	ZBTB16
KDR	NCOA2	POLD1	RNF213	TAF15	ZBTB20
KEAP1	NCOR1	POLE	RNF43	TANGO6	ZBTB7B
KEL	NCOR2	POLQ	RNF6	TAS2R1	ZEB1
KIFC1	NF1	POM121L12	ROBO2	TBX3	ZFHX3
KIT	NF2	POT1	ROS1	TCF4	ZFHX4
KLF4	NFE2L2	POU2F2	RPL22	TCF7L2	ZFP36L1
KLF5	NFKB2	PPM1D	RPS6KA3	TCHH	ZFX
KLHL36	NFKBIA	PPP3CA	RUNX1	TCIRG1	ZIC4
KLHL6	NFKBIE	PPP6C	RUNX1T1	TCL1A	ZMYM3
KMT2A	NHLRC1	PPT2	RYR1	TET1	ZNF148
KMT2B	NIN	PRAMEF12	RYR2	TET2	ZNF165
KMT2C	NIPBL	PRB1	SALL4	TG	ZNF429
KMT2D	NKX2-1	PRB2	SATB1	TGFBR2	ZNF521
KRT15	NONO	PRB3	SCN11A	TGIF1	ZNF679
KRT38	NOTCH1	PRDM1	SCN2A	TLL1	ZNF680
KRTAP9-1	NOTCH2	PRDM2	SCN7A	TMEM30A	ZNF716
LAMA5	NPFFR2	PRDM8	SCN9A	TMEM51	ZNF717
LATS1	NPRL2	PREX2	SDC4	TMSB4X	ZNF721
LATS2	NRAS	PRF1	SELP	TMTC1	ZNF814

LDB1	NRK	PRKAB2	SEMA3G	TNFAIP3	ZNF98
LOX	NSD1	PRKAR1A	SET	TNFRSF14	ZNRF3
LPAR4	NTNG1	PRKCD	SETBP1	TNRC18	ZRSR2
LPP	NTRK3	PRRX1	SETD1B	TNRC6B	ZXDB
LRIG3	NUMA1	PRSS54	SETD2	TOP1	
LRP1B	NUP93	PRSS58	SETDB1	TP53	
LRRK2	NXF1	PSIP1	SF3B1	TP63	
LRRN3	OBSCN	PTCH1	SGK1	TPCN1	

Table S5-2. List of same-lineage tumour suppressor genes identified.

BRCA2	TTN	PLXNB2	NFKB2
ANK1	LRRK2	SYNE1	BCL11B
UBR5	BIRC6	UBR5	DNAH9
UBR5	ZFHX3	APOB	MUC16