Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in (34^{th}cycle)

Algorithms for cancer genome data analysis

Learning techniques for ITH modeling and gene fusion classification

By

Marilisa Montemurro

******

Supervisor(s):
Prof.ssa Elisa Ficarra, Supervisor

Doctoral Examination Committee:
Prof. Andrea Acquaviva. , Referee, Università degli Studi di Bologna, Italy
Dr.ssa Loredana Martignetti, Referee, Institut Curie, Paris, France
Dr.ssa Kaisa Huhtinen, Institute of Biomedicine, Turku, Finland
Dr.ssa Elisabetta Farella, Fondazione Bruno Kessler, Povo (TN), Italy
Prof. Alfredo Benso, Politecnico di Torino, Italy

Politecnico di Torino
2022

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Marilisa Montemurro
2022

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

A mia madre e mio padre, mio fratello Giovanni, il mio amore Claudio e il dolce gatto Trent

Acknowledgements

Reaching the end of my Ph.D. program represents an invaluable achievement. I have worked a lot, but without the supervision and the support of many people, this accomplishment would not have been possible. First and foremost, I want to express my gratitude to my supervisors, Prof. Elisa Ficarra and Prof. Andrea Bertotti, for their invaluable advice, unwavering support, and patience throughout my Ph.D. program. Their knowledge and experience have inspired me throughout my academic research.

I would also like to thank Dr. Elena Grassi and Dr. Gianvito Urgese for their technical and personal support. They have been my mentors and, when needed, caring confidants.

I would like to thank Dr. Marta Lovino for her cooperation and treasured support. She has been an invaluable ally.

I would also like to thank Marco Viviani, who, together with Dr. Elena Grassi, contributed to bringing funny, humor, and light-heartedness to this challenging period.

Additionally, I want to thank all my co-authors and everyone I have had the chance to meet and work with during my Ph.D. program. All of them have enriched me and gifted me with some precious teaching.

Finally, I would like to express my gratitude to my parents, my brother, my fiancé Claudio, and my friends. Without their precious understanding, love, and encouragement over the past few years, it would have been impossible for me to achieve this goal.

Abstract

The introduction of next-generation sequencing (NGS) technology resulted in an explosion of genomic sequencing data. To extract new and useful knowledge, new computational strategies for managing and investigating such data are required.

The first part of this thesis is dedicated to computational methods developed to model intra-tumor heterogeneity. Cancer is an evolving entity, and the evolutionary properties of each tumor are likely to play a critical role in shaping its natural behavior and how it responds to therapy. In fact, during the evolution of the disease, cancer cells differentiate, giving birth to subpopulations (subclones) characterized by a distinguishable set of mutations. This phenomenon, known as intra-tumor heterogeneity (ITH), may be studied using Copy Number Aberrations (CNAs). Nowadays, ITH can be assessed at the highest possible resolution using single-cell DNA (scDNA) sequencing technology. However, since the technology required to generate large scDNA sequencing datasets is relatively recent, dedicated analytical approaches are still lacking. The first part of this Ph.D. thesis has been dedicated to designing new computational methods based on statistical and machine learning techniques to manage scDNA data and unveil spatial ITH.

- In this context, a tool capable of producing multi-sample CNA analysis on large-scale scDNA sequencing data and investigating spatial and temporal tumor heterogeneity has been developed. The main methodological contribution has been leveraging the advantages the of existing approaches, through a different, completely open, pipeline which, for the first time, integrates scCNA data from multiple samples to start investigating ITH from a qualitative point of view.

- Secondly, a study on clustering methods applied to scCNA data is presented. Clustering methods are increasingly applied to scDNA sequencing data to infer

the subclonal structure of cancer. However, the complexity of these data exacerbates some data-science issues and affects clustering results. Additionally, determining whether such inferences are accurate and clusters recapitulate the actual cell phylogeny is not trivial, mainly because ground truth information is unavailable for most experimental settings. Here, by exploiting simulated sequencing data representing known phylogenies of cancer cells, a formal and systematic assessment of well-known clustering methods is presented to study their performance and identify the approach providing the most accurate reconstruction of phylogenetic relationships.

- Finally, a tool to explore the extent of spatial heterogeneity in multi-regional tumor sampling is proposed. The spatial distribution of subclones within a tumor mass can, in principle, be studied using scCNA profiles from multiple samples of the same tumor. However, the existing methods for scCNA analysis are still limited. Many of them only identify the total copy-number, while a few infer the tumor phylogeny using the computed CNAs. An instrument capable of exploiting both the granularity of single-cell DNA data and multi-sample analysis to quantify ITH still does not exist. For this reason, PhyliCS has been developed.

  PhyliCS is the first tool that exploits scCNA data from multiple samples from the same tumor to estimate whether the different clones of a tumor are well mixed or spatially separated.

  In this regard, the SHscore (Spatial Heterogeneity score) is the key methodological contribution. It is a novel metric that allows to quantify how far cells from various samples from the same patient have diverged in their CN landscapes. The SHscore has been evaluated in a variety of simulation settings. Results show that the proposed score accurately represents heterogeneity in the clonal structure of multiple samples and indirectly reflects the evolutionary history of tumor subsamples.

Given the significant contribution of AI techniques in the study of complex biological phenomena characterized by a lack of domain understanding, they were adopted to investigate the oncogenic potential of gene fusions. Gene fusions are one of the most common somatic mutations and are considered to be responsible for 20% of global human cancer morbidity. However, not all gene fusions are oncogenic.

Indeed, some are genuinely expressed in normal human cells or constitute passenger events. Nevertheless, the biological mechanisms which lead from gene fusions to tumorigenesis are not fully understood, and theoretical formulations of this complex phenomenon are still lacking. Therefore, AI algorithms represent an opportunity to infer the causal links between gene fusions and carcinogenesis directly from data. The second part of this thesis has been devoted to the application of deep-learning techniques to the complex task of classifying gene fusions as oncogenic or not oncogenic.

- In this context, a tool based on a specifically designed neural network has been proposed to classify gene fusions as oncogenic or not oncogenic. Identifying potentially oncogenic gene fusions may improve affected patients' diagnosis and treatment. Previous approaches to this issue exploited protein domains, specific gene-related information, to predict the oncogenic potential of the gene functions. The proposed model profits from the earlier findings and includes the microRNAs in the oncogenic assessment. Specifically, the designed neural network integrates information related to transcription factors, gene ontologies, microRNAs, and other detailed information related to the functions of the genes involved in the fusion and the gene fusion structure. The designed neural network outperformed state-of-the-art tools.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of next-generation sequencing (NGS) technology has enabled extensive genome sequencing, revolutionizing medical research. NGS allows to extract millions of sequences, in a single experiment, and use them to analyze relationships, compute statistics, and model complex biological phenomena with very high accuracy. As a result, the amount of data shared to public and private databases is continuously increasing, representing an invaluable source of new knowledge on the human genome [1]. The ability to interpret such data has the potential to help develop effective diagnostic and predictive tools in the treatment of cancer and other complex diseases.

However, data produced from complex systems are complex themselves. Sequencing data are inherently sparse, noisy, and high-dimensional, and, for this reason, distinguishing between the relevant interactions in biological networks and those due to chance is not a trivial task [2].

In this context, Big Data analytics, which allows uncovering hidden patterns and unknown correlations from large-scale datasets, represents a valuable instrument [3].

In this thesis, machine and deep learning techniques have been applied to human DNA and RNA sequences to develop instruments to tackle human cancer data complexity.

The first part of this thesis is dedicated to developing computational methods to characterize intra-tumor heterogeneity (ITH). ITH refers to the coexistence of genetically different cancer cell subpopulations (clones or subclones) within the same

tumor. The clonal composition of each tumor is likely to play a critical role in shaping its natural behavior and how it responds to therapy. Different subclones within the same tumor adapt differently to the external environment and, most importantly, react differently to treatments. For this reason, studying the clonal structure of tumors is one of the most crucial tasks in cancer research nowadays.

Thanks to the emerging large-scale single-cell DNA (scDNA) sequencing technology which allows the extraction of the genomic profile of thousands of individual cells, ITH can be assessed at the highest possible resolution. However, since this technology is quite recent, dedicated analytical approaches are still lacking.

The main methodological contribution of this part of the thesis has been proposing methods and applications to assess ITH exploiting copy-number aberration (CNA) profiles computed on scDNA sequencing data. Specifically, a pipeline to manage large-scale single-cell CNA (scCNA) data, produced by third-party tools, has been proposed; the first formal and systematic assessment on clustering techniques applied to scCNA data has been performed; a tool and a score for the quantification of spatial ITH exploiting scCNA data from multiple tumor samples has been developed.

Given the significant impact of AI-techniques in the study of complex biological phenomena characterized by a poor domain understanding, they were adopted to investigate the oncogenic potential of gene fusions. To this purpose is dedicated the second part of this thesis. Gene fusions are a somatic mutation in which two genes break and, erroneously, fuse together to give birth to a hybrid gene that may be responsible for oncogenesis, tumor development, and poor treatment response. However, not all gene fusions are oncogenic, and correctly identifying them may improve affected patients' diagnosis and treatment.

Unfortunately, the biological mechanisms underlying tumorigenesis initiated by gene fusions are not fully understood, and theoretical formulations of this complex phenomenon are still lacking. Therefore, AI algorithms offer the opportunity to infer the causal links between gene fusions and cancer, exploiting the available abundance of data.

In this regard, the main methodological contribution has been the creation of a tool that integrates structural and functional features of the fused genes with the information about some regulatory and post-regulatory processes. These features are used to classify gene fusions as oncogenic or not oncogenic, exploiting an ad-hoc designed MLP-based architecture.

The thesis is organized as follows.

Chapter 2 will provide biological background and the related computational problems assessed by this thesis.

Chapters 3, 4 and 5 will present the methods developed to address the problem of ITH characterization and their results.

Chapter 6 will describe the model designed to identify the oncogenic potential of gene fusions.

In the end, Chapter 7 will provide some conclusions, highlighting the key strengths of the proposed methods.

Chapter 2

Background

Cancer is a disease of the genome that arises from the accumulation of mutations during cell life-cycle [4]. Despite the decreasing trend in cancer mortality, it remains a severe public health problem worldwide and is the second leading cause of death in the United States [5]. Nowadays, Bioinformatics is providing the instruments to speed-up cancer research.

2.1 NGS: genomics and data science

One of the most significant steps forwards in finding new diagnostic and therapeutic tools has been the advent of *next-generation sequencing (NGS)* technologies in 2006. NGS is a high-throughput sequencing technology that allows to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA. By comparing such nucleotide sequences with a reference genome or transcriptome, it is possible to spot alterations with respect what is considered the normal structure of the sequenced area and, eventually, make hypotheses on the anomalies which may have caused a given disease, including cancer [6].

In particular, NGS promoted the advent of precision medicine (PM). In fact, the possibility of massively sequencing patient's genomes at a relatively low cost has enabled the identification of individual tumor mutations (biomarkers) that allow to predict its evolution and response to therapies. Therefore, it is possible to design specific therapies for individual patients [7].

However, NGS poses some technical challenges which require computational frameworks to be assessed.

2.1.1 The rapid growth of sequencing data

NGS technologies represent a source of big data. In fact, the amount of data generated by sequencers is continuously increasing thanks to the technical advancements and the lowering of sequencing costs.

The National Human Genome Research Institute (NHGRI) has been tracking the costs associated with DNA sequencing performed at the sequencing centers funded by the Institute since 2001 [8]. According to their results, the costs of sequencing an entire human genome dropped from $95,263,072 in 2001 to $562 in 2021, with a considerable decrease in 2008, when the sequencing costs fell from $7,147,571 to $342,502 in a single year.

The advances in sequencing technology and the evolution of computational frameworks are speeding up both data generation and data sharing. In the future, it is estimated that the sequencing of an entire human genome may take less than 24h, dropping from the 2 to 8 weeks required with the current technology, and from the 13 years required to sequence the first human genome by the Human Genome Project [9].

Besides, thanks to the continuously emerging sequencing protocols, also the variety of genomic data is increasing over the years (e.g, WGS [10], ChiP-seq [11], RNA-seq [12], ATAC-seq [13], sc-seq [14], etc.).

As a result, the volume of genomic data grows more than that of many other data-intensive disciplines (e.g., astronomy, sociology, TOP 500 supercomputers, IP traffic) [15]. Researchers believe that if this data generation growth trend remains constant, soon genomics applications will generate more data than social media [16]. In fact, it is estimated that genomics projects will generate 40 exabytes of data by 2025 [17].

2.1.2 AI for genomics

To tackle the growing volume and complexity of genomic data, in the last years, researchers have been moving to Artificial Intelligence to identify meaningful patterns and extract new biological knowledge.

AI overview

The expression *Artificial Intelligence (AI)* was coined by John McCarthy during the Dartmouth Artificial Intelligence conference held in 1956. It was chosen as an umbrella expression, comprehensive of the many different fields discussed during the conference.

Nowadays, the term AI refers to the set of devices, algorithms, and applications that allow computers to mimic human intelligence [18]. In particular, AI provides the instruments to uncover hidden patterns and unknown correlations from large-scale and complex datasets. The field of AI is one of the most dynamic ones and is rapidly evolving.

Machine learning (ML) and *deep learning (DL)* are two macro-areas of application in AI. In particular, machine learning refers to the capability of a machine to learn from a large dataset without being programmed on what it learns. Deep learning, instead, is a relatively modern computational paradigm that allows to learn patterns in the data imitating the working principles of human neurons, using layered networks of computational units (the *artificial neurons*) named *artificial neural networks*.

Learning may be supervised (when algorithms are fed with labeled training data representing the categories which the model should learn) or unsupervised (when the algorithms can recognize patterns in the datasets without any hint from humans). AI algorithms are, then, used to make informed decisions using what they have learned.

The great advantage of AI algorithms is their ability to find linear and non-linear correlations in huge datasets made of complex data, with an accuracy and a velocity that would be impossible to reach by any human being. This is the reason why life scientists need AI-based tools to profit from the vast amount of genomic data at their disposal.

AI applications in genomics

Although AI techniques are at their early stages in genomics, some applications do exist. For example, they have been used to identify genetic disorders from people's faces [19], to discriminate between disease-causing genomic variants and benign variants [20], to identify the primary type of cancer from a liquid biopsy [21], to predict the evolution of a specific type of tumor in a patient [22].

2.1.3 AI techniques in cancer research

Cancer genomics is one of the research areas empowered by the explosion of AI applications.

Many AI-based models have been proposed to improve variant discovery and classification accuracy, to classify tumors into subtypes, to predict the disease outcome and response to therapies, to incorporate high-dimensional data sets, and to integrate multi-omics [23–31].

Many reviews have focused on AI applications for cancer study [32–35]. This thesis will deal with applications for intra-tumor heterogeneity (ITH) characterisation, using single-cell copy-number aberration (scCNA) data, and oncogenic gene fusion prioritization.

Algorithms for ITH characterization

Tumors are caused by the accumulation of somatic mutations. The set of mutations collected by the tumor founder cell, known as clonal, is inherited by the entire progeny of the tumor. Mutations that arise in an already existing tumor are only passed on to subpopulations of cells and are referred to as subclonal [36, 37]. As a result, cancer cells exhibit intrinsic genetic diversity, known as *intra-tumor heterogeneity (ITH)* [38] which is recognized as a major cause of tumor recurrence, and treatment failure [39, 38, 40–42].

Regionally separated heterogeneous clones can lead to sampling bias. For instance, there is increasing evidence that resistance to some targeted cancer treatments may result from the expansion of preexisting low-frequency tumor cell populations harboring somatic mutations that provide resistance to the targeted drug [43]. Such

low-frequency populations may not be detected by single-tumor biopsy, affecting the identification of tumor biomarkers. To this regard, many studies have shown that using multiple samples taken from different regions of the same tumor improves the ability to infer the subclonal structure of tumors and to identify the most appropriate drugs [44–47, 38–40, 48].

Computational frameworks currently used to characterize ITH at the genome level are mainly based on statistical methods and clustering techniques.

The most common method for assessing ITH is to use deconvolution techniques on bulk DNA sequencing data [49, 50]. Bulk DNA sequencing allows to sequence a mix of millions of cells and generate a mixed genomic profile of all subclones. It is considered the standard tool to generate genomic data at the cell population level, so its results are considered reliable and stable; however, the main drawback is that the subclonal tumor structure is hidden, so bioinformatics techniques are required to infer it. In detail, the generated sequences are first aligned to a reference genome and then processed with mutation callers [51] to identify the somatic mutations carried by the tumor sample. The *variant allele frequency (VAF)* for each somatic mutation can be calculated by dividing the number of sequence reads matching that variant by the read coverage at that locus. An estimate of the fraction of tumor cells carrying each mutation can be obtained using mutation allele frequencies and accounting for copy number variations. A set of mutations can then be used as a marker for a population of cells, allowing estimation of the fraction of tumor cells belonging to the corresponding subclone. Clustering algorithms can be used to determine the cancer cell fractions of each subclone and, ultimately, the tumor evolutionary tree (*phylogeny*) [52–61].

Bulk DNA data deconvolution techniques mainly rely on machine learning models and statistical computations to infer tumor subclonality indirectly, frequently resulting in an ensemble view dominated by the prevalent clones. Nowadays, emerging single-cell DNA sequencing (scDNA-seq) technologies offer an extraordinary opportunity to tackle such issues, as they allow to sequence the genome of individual cells and study tumor heterogeneity with unprecedented resolution.

In particular, single-cell low-coverage whole-genome sequencing is suited for detecting copy-number aberrations (CNAs), which can be exploited to reconstruct cell population subclonal structure using clustering techniques [62]. CNAs are a common type of somatic mutation, consisting of the alteration of the expected

number of copies of one or more regions of the genome (the normal copy-number for a diploid genome, like the human one, is 2) [63].

The majority of existing methods for single-cell CNA (scCNA) data [64–71] only identify the total copy-number. A few of them also infer the tumor phylogeny applying clustering algorithms to the CNAs they computed [72].

Even if it is out of the scope of this thesis, it is worth mentioning that some studies have proposed methods to evaluate ITH using gene expression [73–75] or protein networks [76]. Other studies presented imaging techniques based on neural networks such as CNN's and ResNets [77] to visualize tumor heterogeneity from a morphological perspective.

The first part of this thesis, instead, will focus on ML techniques applied to scCNA data to characterize and quantify ITH from the genomic point of view.

Algorithms for gene fusion classification

Gene fusions are a common type of mutation that results from the joining of two independent genes [78]. In the most common case, the gene at the 5' retains the promoter and the 5' UTR region while the 3' gene retains its end sequence in the fusion. Gene fusions are considered to be responsible for 20% of cancer-related deaths.

The first fusion gene, the Philadelphia chromosome [79], was identified in chronic myelogenous leukemia in 1973 [80]. After that finding, other fusion genes have been associated with various hematological cancers [81–84] and, recently, discovered in different solid tumors, including sarcomas, carcinomas, and central nervous system tumors [85–88].

The discovery of many cancer-related gene fusions has impacted clinal care. They are used to diagnose a variety of cancers [89, 90], identify molecular cancer subtypes [91, 92], stratify patients [93, 94], monitor residual disease after treatment [95, 96], and predict relapse [96]. Notably, fusion transcripts are also promising therapeutic targets [97–99], with the potential to improve patient outcomes significantly.

However, not all gene fusions are oncogenic. Some of them are expressed in normal tissues [100] while others are discarded by DNA repair mechanisms [101].

The first step to identify oncogenic gene fusions is analyzing RNA sequencing data, using one of the existing fusion detection tools (e.g., ChimeraScan [102], deFuse [103], STARfusion [104], FusionCatcher [105], TopHat-Fusion [106], SOAP-fusion [107], etc.). Such methods generate a long list of putative gene fusions, most of which are false positives. Importantly, these tools do not provide any statement about gene fusions' involvement in tumorigenesis. Hence, prioritizers are gaining popularity: these are tools that can automatically recognize which of the many gene fusions carry an oncogenic potential and, therefore, are the best candidates to be analyzed in the laboratory.

From a computational point of view, however, the problem of identifying which gene fusions are actually oncogenic is challenging and anything but simple. This problem was addressed from 2015 onwards when some prioritizers were released in the literature. Among the firsts, Oncofuse and Pegasus [108, 109] are to be mentioned: they exploit general contextual information related to genes (in particular protein domains) to establish through Decision Trees and Support Vector Machines whether the new fusions are oncogenic or not. However, both models achieve precision and recall of no more than 70%. More recently, DEEPrior [110] has been proposed. It is a prioritizer for the search for oncogenic fusions, based on CNN and LSTM architectures, that directly exploits the resulting protein sequence to determine whether the fusion is oncogenic or not. In this case, the recall of prioritized gene fusions is about 80%.

The computational complexity of this biological problem lies in the fact that the oncogenicity property is very complex and determined by a set of not fully known factors within the fused gene and the molecules involved in its regulation and expression. Indeed, there is no set of genes, protein domains, or transcription factors that alone can effectively predict the oncogenic potential of gene fusions.

In addition, the task is complicated by the need to measure the results on a set of gene fusions whose genes have never been used in the training phase. Such an imposition is necessary as all databases relating to gene fusions cover only a tiny portion of all possible existing gene fusions. Therefore, it is advisable to have models that can sufficiently generalize the concepts learned on the training process.

Learning algorithms seem to be a valid technological solution to address this problem. In fact, they have shown brilliant results when employed to examine complex phenomena with poorly understood properties. Deep learning, in particular,

has the incredible capacity to learn, on its own, to describe phenomena as a tiered hierarchy of concepts and finding hidden patterns among them. Because they can learn directly from data, such approaches are highly beneficial when used to address problems characterized by a lack of domain understanding of their features. Furthermore, DL algorithms allow to avoid human intervention and feature engineering, particularly when data size is huge.

Therefore, in the second part of this thesis, DL techniques have been used to address the complex task of prioritizing gene fusions. The result has been an application, based on an ad-hoc designed MLP-based architecture, that integrates structural and functional features of the fused genes with the information about some regulatory and post-regulatory processes to predict the oncogenic potential of gene-fusions. Indeed, it is still unclear which features are the most relevant and how they interact to enable oncogenic functional activities. The MLP-based model is then used to learn through the data these interdependent relationships.

Part I

Intra-tumor Heterogeneity Characterization

Chapter 3

Single-Cell Dna Sequencing Data: A Pipeline For Multi-Sample Analysis

This chapter starts investigating how tumor spatial heterogeneity can be addressed with large-scale scCNA datasets.

Nowadays, single-cell DNA (sc-DNA) sequencing is showing up to be a valuable instrument to investigate intra and inter-tumor heterogeneity and infer its evolutionary dynamics, by using the high-resolution data it produces. That is why the demand for analytical tools to manage this kind of data is increasing.

In this context, a pipeline capable of producing multi-sample copy-number aberrations (CNA) analysis on large-scale single-cell DNA sequencing data and investigate spatial and temporal tumor heterogeneity is proposed.

3.1 Scientific Background

One of the main challenges for cancer researchers is understanding the evolutionary dynamics of the disease. In fact, it is largely known that cancer is an evolving entity and the evolutionary properties of each tumor are likely to play a critical role in shaping its natural behavior and how it responds to therapy [111].

Emerging single-cell DNA sequencing technologies allow profiling individual cells, highlighting differences among them, and assessing tumor heterogeneity with an unprecedented detail level. Additionally, it is possible to adopt phylogenetic

reconstruction techniques to infer the evolutionary history of cell sub-populations within single-cancer cases. One way of exploiting single-cell DNA sequencing to address tumor heterogeneity and evolution investigation is performing copy-number aberration () analysis [112].

At the time of writing, single-cell DNA sequencing usually requires isolating cells and performing whole-genome sequencing (WGS) on each cell; as a result, only a few of them can be analyzed together in one experiment, making heterogeneity studies less effective. A few technological solutions for large-scale single-cell DNA sequencing have been proposed [113, 114]. The most popular one is *10x Genomics* technology [114]. The company offers a new system, named *Chromium System*, which is able of partitioning and barcoding hundreds to thousands of individual cells into GEMs: the DNA in each GEM can then be fragmented and amplified into short-read libraries suitable for Illumina sequencing, with each fragment receiving a 14-base molecular barcode unique to its GEM of origin. Additionally, 10x Genomics provides a proprietary pipeline, *Cell Ranger DNA*, that identifies events (down to 2Mbp for a single cell and down to 200kbp for clusters of cells) and infers a phylogenetic tree, which can be explored, together with its associated heatmap, using an interactive browser, *Loupe scDNA Browser*. Cell Ranger DNA, right now, allows to execute only single-sample analysis and, being a closed platform, it cannot be customized.

In addition to 10x Genomics pipeline, another pair of tools, capable of performing calling on sc-DNA data, exist: *Ginkgo* [67] and *SCCNV* [115]. For this project, as an alternative to Cell Ranger DNA, only Ginkgo has been considered, which is more diffused and has already been considered as a reference tool for this kind of application [116]. Ginkgo is a freely available open-source web-based application for copy-number variants' automated and interactive analysis. Ginkgo allows to upload single-cell alignment files (one file for each cell) to define different parameters and, after calling, to analyze the results using a web interface. Additionally, it computes and shows a phylogenetic tree, which can be computed using different distance metrics and clustering algorithms, and draws some heatmaps showing clusters of clones. The main Ginkgo drawback is that being a web application, it is not easy to embed it into an automated pipeline and run it on a cluster, like an HPC platform; moreover, it suffers from the delays introduced by the network.

Additionally, as far as we know, it has been validated on datasets significantly smaller than the ones managed by 10x software.

Here a different, completely open, pipeline is proposed: it combines the advantage of transparently dealing with 10x sequencing data of Cell Ranger DNA, with the openness and flexibility of Ginkgo, used in a stand-alone fashion, in order to perform a *multi-sample* single-cell analysis, on *large-scale* datasets. The pipeline is freely available at the following link: https://github.com/vodkatad/biloba.

3.2 Materials and Methods

The proposed pipeline is organized in two main workflows: the first one manages data preparation, processing, and validation for each of the tumor samples of interest; the second one aggregates the results of the preceding phase to perform multi-sample analysis and evaluate inter and intra-sample heterogeneity. The pipeline is automatically handled by Snakemake [117] and comprises a stand-alone version of Ginkgo, deprived of the web utilities, which were not helpful for a command-line tool. The newly implemented modules have been written in Python and C++.

3.2.1 Single-sample analysis

Data preparation. The pipeline starts with the execution of Cell Ranger DNA. The tool demultiplexes BCL files obtained from Chromium-prepared sequencing samples and produces FASTQ files where each read contains the barcode corresponding to the cell/GEM which originated it. After that, it performs reference alignment, filtering out poor quality mappings (MAPQ < 30), calls CNA events, and performs hierarchical clustering of cells on the base of their CNA profiles. Additionally, when the CN of a cell varies too often and too much in adjacent bins or when the algorithm which inferred its ploidy emitted a low confidence score, the cell is marked as noisy (please refer to 10x Genomics documentation for a detailed explanation). 10x pipeline is a closed software, so it is impossible to customize the analysis. This goal is reached by re-analyzing data with Ginkgo which internal details and intermediate results are not hidden to the user. The two platforms have been integrated using a module in charge of manipulating Cell Ranger DNA results so that they can be used as input files for Ginkgo. First of all, a *demultiplexer*, implemented by using functions implemented

in the SeqAn C++ template library [118], splits the alignment file produced by Cell Ranger into separate alignment files for the different cells. After that, a quality filter is applied to filter out multi-mapping, and poor quality reads (MAPQ$<$ 30). At this point, data are ready to be provided as input files to Ginkgo, which performs CNA calling on single-cell aligned data.

Data post-processing. Ginkgo CNA profiles, organized in a matrix, are used to compute hierarchical clustering using the Euclidean distance and the complete linkage method. Clustering results are used to draw a heatmap with its associated dendrogram: in this way, it is possible to compare the results of the two tools and check if they are consistent. Moreover, in order to provide functional annotations, CNA events are annotated with the corresponding gene symbols.

Finally, the user can filter out cells considered insignificant for the following analysis. For example, it is possible to select a range of accepted mean ploidies to filter out diploid cells and remove the immune, and stromal normal infiltrate from tumor biopsies.

3.2.2 Multi-sample analysis

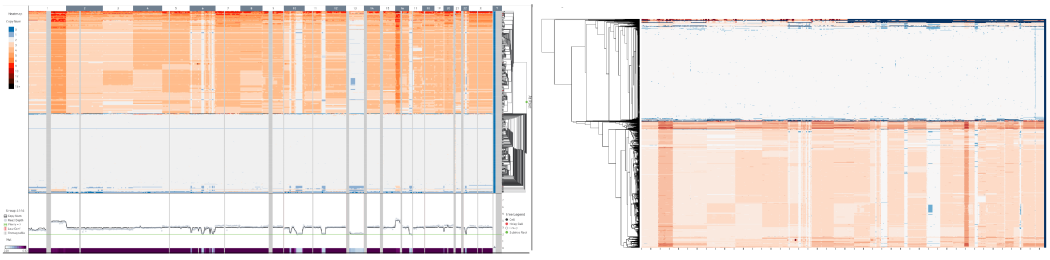
CNA calls from multiple data are merged and clustered together to produce an aggregated dendrogram with its heatmap, where the cells coming from different samples can be identified using different colored labels. The user can now evaluate, in a qualitative way, how much the different samples are evolutionary distant: cells whose profiles are very similar, because they went through a common mutational pattern, carry on a similar genetic signal, and this will be evident from the heatmap, which will show a pretty consistent trend. Additionally, the clustering algorithm will mix up cells, in this case, producing a dendrogram where the leaves, corresponding to the cells from the different samples, will not segregate in blocks. In the opposite case, if the two samples are very different (e.g., spatial segregation of clones happened, so two samples originated by the same original cancer tissue are very different), this will affect the aggregated results in a way that both the tree and the heatmap will show well-separated blocks corresponding to the different samples.

3.3 Results

The pipeline was tested on two single-cell CNA datasets produced by Cell Ranger DNA and published on the 10x Genomics website. The datasets derive from two sections (`section_A`, `section_B`) of the same frozen breast tumor tissue, from a triple negative ductal carcinoma, and contain the data of 2137 and 2224 cells, respectively.

3.3.1 Single-Sample Analysis

Here, the results obtained executing the first part of our pipeline on `section_B` dataset are described.
Figure 3.1 compares the results of Cell Ranger DNA execution to the ones obtained by applying the first part of our pipeline (demultiplexer, filter, and Ginkgo), condensed in the form of two heatmaps. Specifically, Cell Ranger DNA heatmap can be observed only through their browser and cannot show more than 512 cells; this means that in this case, cells with similar CNA profiles are grouped and shown together. Anyhow, as it can be seen, the two methodologies obtain the same results: two distinct cell sub-populations are found, one having a mean copy-number ranging from 3 to 4, and one diploid cell sub-population, probably corresponding to normal tissue infiltrate. Regarding tumor cells, the two heatmaps show a very similar trend, with the same amplified and deleted regions.



(a) Cell Ranger DNA heatmap

(b) Our heatmap

Fig. 3.1 Comparison between the heatmaps obtained by executing hierarchical clustering on CNA profiles by Cell Ranger DNA (3.1a) and our pipeline (3.1b). Blu areas represent deletions, red areas represent amplifications and white areas represent diploid segments: big CN losses/gains are indicated by a darker blue/red.

Additionally, the Spearman correlation coefficient was computed between each pair of tumoral CNA profiles, produced by the two procedures, for the same cell (cells with mean ploidy ranging from 1.5 to 2.7 have been discarded for this analysis). This correlation (mean value $\cong 0.76$) further supports the impression derived from the heatmaps: the two pipelines produce superimposable results (Figure 3.2).

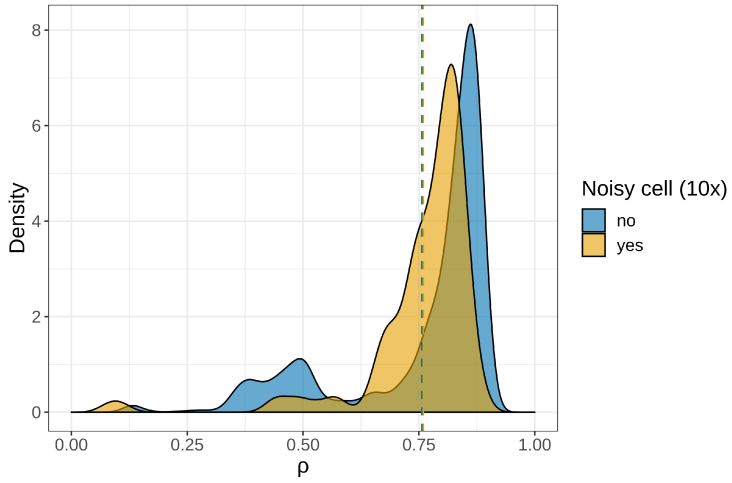


Fig. 3.2 Distribution of the Spearman correlation coefficients computed for each pair of tumoral CNA profiles produced by Cell Ranger DNA and our pipeline. 10x noisy cells have appear in a separate plot (yellow) to clearly show that also in this case results correlate.

3.3.2 Multi-Sample Analysis

Before going on with the multi-sample analysis, cells with a mean ploidy ranging from 1.5 to 2.7 were filtered out, assuming they were derived from normal tissue infiltrate. This resulted in a set of 117 tumoral cells from *section_A* and 1257 tumoral cells from *section_B*; *section_A* analysis results showed that normal and tumoral cell numbers were unbalanced, in this case, with a high prevalence of diploid cells. Figure 3.3 shows the result of the hierarchical clustering performed on the aggregated results for the two cancer sections. The cells from the two sections, indicated by two different colors in the vertical bar between the dendrogram and the heatmap, present a similar CNA trend and, for this reason, have been homogeneously mixed up by the phylogenetic reconstruction algorithm. This is consistent with the hypothesis that the two samples, derived from the same tumor, share their mutational history.

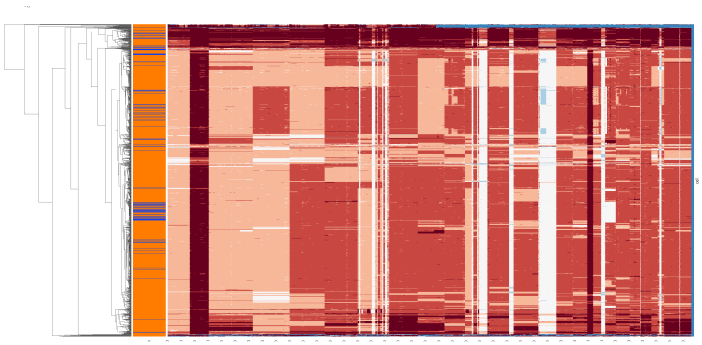


Fig. 3.3 Multi-sample phylogenetic tree and heatmap: blue bars correspond to `section_A` cells and orange bars correspond to `section_B` cells.

Finally, to provide robustness to the analysis, the multi-sample analysis was performed on two independent and completely separate single-cell datasets used to validate Ginkgo by its authors. Both of them come from already published studies, analyzing and highly heterogeneous triple-negative ductal carcinoma breast cancer [112] and circulating tumor cells from lung adenocarcinomas [119], respectively. In this case, diploid cells were not removed to let the reader compare the results with those reported by the Ginkgo authors. Figure 3.4 shows that, in this case, the cells cluster into well-separated sub-populations corresponding to the different samples which originated it.

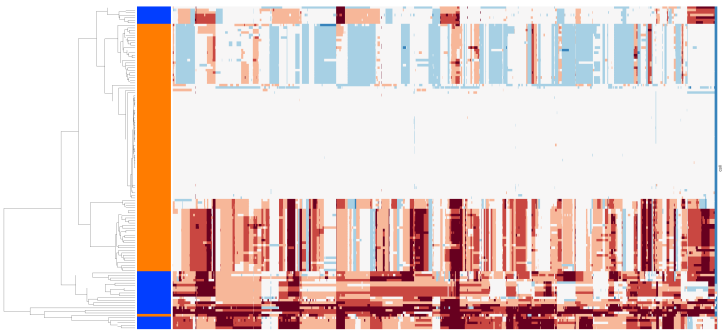


Fig. 3.4 Multi-sample phylogenetic tree and heatmap: blue bars correspond to breast tumor cells and orange bars correspond to lung tumor cells.

3.4 Conclusions

This chapter has presented a tunable pipeline to analyze large-scale scDNA experiments obtained with the 10x technology. This approach overcomes the limits of closed source software, giving on the one hand researchers the possibility to explore the complete breadth of their data and, on the other, a fully-fledged and easy to run pipeline to obtain multi-sample and heterogeneity visualizations. It has been shown that Ginkgo produces results equivalent to Cell Ranger DNA pipelines and started investigating how tumor spatial heterogeneity can be addressed with scCNA datasets.

The development of a fully modular pipeline opens many opportunities for further applications, such as leveraging the *demultiplexer* to obtain single-cell alignments for 10x derived scRNA datasets, to exploit third-party software also for RNAseq, or to analyze "barnyard" experiments in the scDNA context.

Chapter 4

Effective Evaluation of Clustering Algorithms on Single-Cell CNA data

Clustering methods are increasingly applied to single-cell DNA sequencing (scDNAseq) data to infer the subclonal structure of cancer. However, the complexity of these data exacerbates some data-science issues and affects clustering results. Additionally, determining whether such inferences are accurate and clusters recapitulate the actual cell phylogeny is not trivial, mainly because ground truth information is unavailable for most experimental settings.

By exploiting simulated sequencing data representing known phylogenies of cancer cells, this chapter proposes a formal and systematic assessment of well-known clustering methods to study their performance and identify the approach providing the most accurate reconstruction of phylogenetic relationships.

4.1 Scientific background

Cancer cells accumulate genetic alterations at every cell division, including sequence variants and structural variations with gross copy number changes of entire genomic regions (i.e., copy number alterations, CNAs). On these premises, similarities in the genomic structure of individual cancer cells can be exploited to estimate the phylogenetic distance across different cells and consequently infer the subclonal

structure of a tumor. For this reason, scDNAseq is becoming an increasingly popular technique [120, 62].

The most common way of inferring a single-cell CNA (scCNA) phylogeny is by performing hierarchical clustering on the CN profiles [67, 121], assuming that similar cells are very likely to have experienced the same mutational events. However, some biases could affect this kind of approach and vitiate the accuracy of the outcome. Specifically, clustering single-cell data exacerbates some biological data-science issues [2]. Indeed, the increasing number of cells which can be sequenced together expands the space of possible cluster assignments, and determining the most meaningful results is not trivial without knowing the underlying biological truth. Additionally, the high-dimensional nature of such data harbors the "curse of dimensionality" [122]: distance metrics stop behaving as expected based on our low-dimensional intuition, and clustering algorithms fail in determining the distance between points. Moreover, the infinite-sites model does not apply to cancer CNAs [123], which intrinsically diminish the power of exploiting similarities in the genomic structure to predict phylogenies.

Although some of these issues have been partially addressed in the context of single-cell RNA methodology [124], in the case of scDNAseq, the extent of available data is still limited, and there is a need for the development of dedicated data analysis methods.

On these premises, the present work aims at proposing a first formal and systematic performance evaluation of nine well-known clustering methods on scCNA data.

A synthetic scCNA dataset has been generated to evaluate the accuracy, stability, run time, and scalability of nine clustering methods. Moreover, the performance of the algorithms has been compared following different pre-processing steps. Finally, we tested the best-performing methods on a real scCNA dataset obtained from colorectal cancer cells. The code used to perform our analysis is available at https://github.com/mmontemurro/clustering_benchmarking.

4.2 Materials and Methods

In the following, we will describe the procedure to generate the simulated and the real scCNA datasets and the evaluation methods we have used for this work.

4.2.1 Simulations

A simulation experiment was designed to compare the performance of the clustering methods on datasets of three different sizes (100, 200, and 400 cells). Each experiment was iterated 50 times for a total of 150 datasets.

Simulations were performed using the method presented by Fan et al. [125] which generates a phylogenetic tree starting from a reference genome, using a generalization of the Beta-Splitting model [126]. When a new edge enters the tree, some new CNAs are generated by sampling from a Poisson distribution (default $\lambda = 2$). The CNA size is determined by sampling from an exponential distribution (default mean=5Mbp), plus a minimum CNA size (default 2Mbp). The kind of alteration (*gain* vs. *loss*) is decided by a binomial distribution (default $p = 0.5$). If a CN gain is sampled, the number of copies to be gained is determined by a geometric distribution (default $p = 0.5$). If a CN loss is sampled, the whole sequence on that region of the allele is deleted. The allele is chosen by drawing from a binomial distribution (default $p = 0.5$). The chromosome and the starting position of the CNA are sampled from a uniform distribution, bounded between 0 and the genome size. The daughter cell inherits all CNAs from the parent node, in addition to its unique CNAs. In agreement with the finite-site model of CNA evolution, new mutations may occur on already mutated sites. Additionally, to mimic the behavior of punctuated evolution [127], at the edges to the root, whole-chromosome amplifications may occur, in addition to focal CNAs. The probability of a chromosome to be amplified at this step is set using a binomial distribution (default $p = 0.2$). Finally, a given multiplying factor may increase the number of CNAs generated at this step. In the end, the leaves of the generated tree represent the cells sampled from the patient, while the internal nodes represent intermediate CN states, which do not exist anymore.

In order to evaluate the ability of clustering methods to produce groups of cells phylogenetically related, the generated trees were converted into easy-to-be-handled Newick format [128] and a set of clusters, directly from the trees, was identified

to be used as ground truth. The clusters are extracted as proposed from Balaban et al. [129] by solving an optimization problem that, given an arbitrary tree, returns the minimum number of clusters such that the maximum pairwise cophenetic distance between leaves in each cluster is lower than a given threshold. The threshold has been chosen according to the empiric observation that using a value equal to the height of each tree, a set of balanced clusters is obtained.

4.2.2 Clustering algorithms and evaluation methods

Since there is no formal evidence that hierarchical clustering should be preferred to other clustering paradigms in this scenario, six among the mostly used methods were selected, which implementation is available: *Affinity Propagation* [130], *Agglomerative Hierarchical clustering* [131], *Birch* [132], *DBSCAN* [133], *HDBSCAN* [134] and *K-Means* [135]. Additionally, four variants of the agglomerative method [131] were tested: *average linkage*, *complete linkage*, *single linkage* and *ward linkage*.

Each clustering method was applied on every simulated dataset in three different scenarios: (i) without any preprocessing stage; (ii) after low variance feature filtering and PCA-based dimensionality reduction and (iii) after low variance feature filtering and UMAP-based dimensionality reduction.

The whole pipeline is fully automated. The Silhouette score maximization heuristic [136] has been used to determine the cluster number for the algorithms requiring it. Through this, a real-world scenario was simulated, in which the cluster number is not known a priori and must be arbitrarily chosen. For each dataset, the optimal number of PCs has been defined based on a randomization method, as described in Peres-Neto et al. [137]. This method consists in shuffling the dataset many times (default $N\_iter = 50$) and computing the percentage of variance explained by the PCs at every iteration. The significance of each PC is then defined as the probability that the permuted variance is greater than that observed one. Based on this, all the PCs characterized by a p-value equal to or below the threshold significance level (default $\alpha = 0.05$) are considered informative.

For each clustering method, the execution time was measured, and the following indices were computed:

- (*stability*) the *Average Proportion of Non-overlapping* (APN). This score measures the average incoherence between full data clustering and clustering based on data in which one dimension was removed. Values closer to 0 indicate good algorithm stability.

- (*accuracy*) the *Adjusted Rand index* (ARI), the *Adjusted Mutual Information* (AMI), the *V-Measure* (VM) and the *Fowlkes-Mallows Index* (FMI). These indices measure the similarity between the ground truth and clustering results. Values closer to 1 indicate good algorithm accuracy.

4.2.3 Single-cell sequencing

A real dataset has been generated by executing a scDNA-seq experiment on the human non-metastatic colorectal cancer-derived cell-line, SW480.

To this purpose, cells were cultured in L-15 medium supplemented with 10% FBS and 1% penicillin-streptomycin. Nuclei isolation was performed according to 10X Genomics protocol [138]. Briefly, 1 million cells were centrifuged (300 rcf for 5 minutes, at 4°C). Cell membranes were then lysed using a pre-chilled lysis buffer, and nuclei were pelleted by centrifugation (850 rcf for 5 minutes, at 4°C). Supernatant was removed, and nuclei were washed twice in PBS (0.04% BSA). After it, nuclei were counted and re-suspended to a 1000 nuclei/ul concentration. Three thousand nuclei were processed accordingly to manufacturer protocol [139], to generate a barcoded DNA library from each nucleus. After QC check, libraries were sequenced on a Novaseq 6000 S1 flow cell (Illumina).

10X Genomics proprietary pipeline [121], *Cell Ranger DNA* was used to filter-out sequencing noise, align the reads against the GrCh38 reference genome, and assign them to valid cell identifiers. The alignment file was demultiplexed into single-cell .bam files, filtering out poor quality reads ($MAPQ < 30$), multimappers, and secondary alignments. Finally, a customized version of *Ginkgo* [67] was used to extract scCNA profiles. The choice to use Ginkgo to call CNAs was motivated by the need for flexibility not fully provided by Cell Ranger DNA.

The resulting dataset contained 399 scCNA profiles.

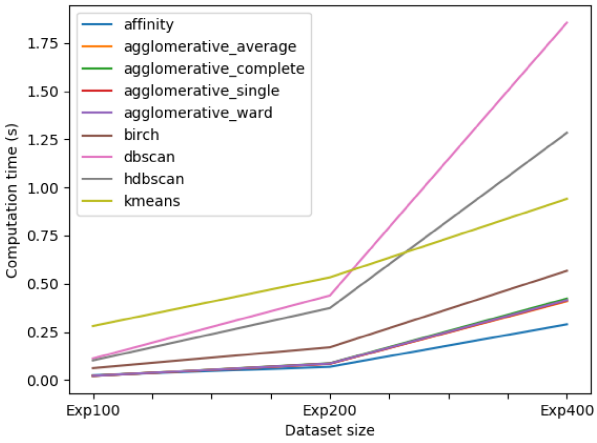


Fig. 4.1 Clustering algorithm evaluation: mean computation time on non-reduced datasets.

4.3 Results and discussion

4.3.1 Evaluating clustering

Each clustering method was applied to every simulated dataset in the three preprocessing scenarios. Therefore, each algorithm was executed 450 times for a total of 4050 clustering results. The evaluation metrics were computed for each algorithm run and then aggregated to summarize the results.

In the following, the main results are summarized.

Computation time

Figure 4.1 shows how the mean computation time increases as the input datasets become larger in the no-preprocessing scenario. When dealing with small datasets, all algorithms achieve comparable performance; as the dataset size increases, density-based algorithms (DBSCAN, HDBSCAN) behave worse. This result was expected since it reflects the complexity of the algorithms.

	Exp100			Exp200			Exp400		
	No prep.	PCA	UMAP	No prep.	PCA	UMAP	No prep.	PCA	UMAP
affinity	0.0	0.06	0.25	0.0	0.05	0.35	0.0	0.05	0.33
aggl. average	0.0	0.02	0.11	0.0	0.02	0.15	0.0	0.01	0.21
aggl. complete	0.0	0.03	0.12	0.0	0.03	0.18	0.0	0.05	0.25
aggl. single	0.0	0.02	0.07	0.0	0.01	0.11	0.0	0.0	0.16
aggl. ward	0.0	0.03	0.12	0.0	0.03	0.17	0.0	0.02	0.24
birch	0.0	0.03	0.1	0.0	0.03	0.16	0.0	0.03	0.21
dbscan	0.0	0.12	0.46	0.09	0.09	0.39	0.04	0.08	0.23
hdbscan	0.0	0.05	0.0	0.0	0.05	0.0	0.0	0.06	0.0
kmeans	0.06	0.03	0.13	0.08	0.05	0.18	0.08	0.08	0.25

Table 4.1 Clustering algorithm evaluation: Mean APN scores.

Stability

Table 4.1 shows the mean APN score over different sizes of the input datasets for the three preprocessing scenarios. All algorithms demonstrated good performance (APN near to 0), in terms of stability, in all tested conditions.

However, in the absence of any preprocessing stage, K-Means and DBSCAN achieve the worse scores. Moreover, all the algorithms were less stable when applied to data preprocessed through PCA or UMAP. This is expected and coherent with the notion that following dimensionality reduction, all the selected features are relevant for classification. As a final remark, it is interesting to notice that increasing the input dataset size from 200 to 400 cells improved the stability of DBSCAN.

Accuracy

Figures 4.2, 4.3 and 4.4 summarize the results of our analysis on clustering accuracy. Algorithms were ranked to identify the most accurate one for each input dataset size and preprocessing scenario. To this purpose, a rank was first assigned to each algorithm based on each validation index, and then the overall performance was computed as the average of the ranks.

The only algorithm which demonstrated good accuracy even in the absence of data preprocessing is Affinity Propagation (AP) clustering. This is reasonable since the AP algorithm was already shown to perform well in various data-science fields, dealing with various kinds of high-dimensional data [140–143]. The reason for the

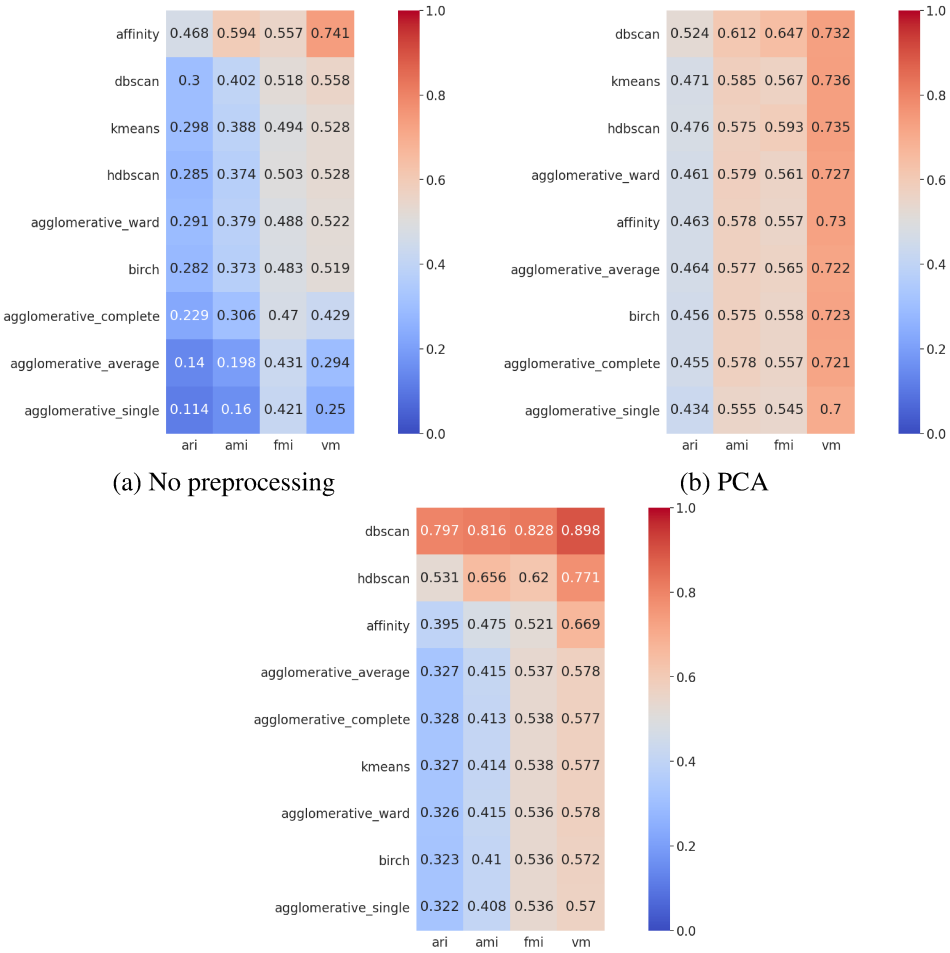
good performance of AP is likely related to the fact that it does not take random samples for cluster centers but considers all points as possible exemplars [144].

On the contrary, it is interesting to notice that Agglomerative clustering based on single and average linkage consistently performed worse than the others, possibly because they are susceptible to noise and, as a consequence, tend to produce a high number of tiny singleton clusters. In contrast, Agglomerative clustering with ward linkage performed better, in accordance with the notion that it generally produces more balanced clusters, and should be preferred when performing hierarchical clustering on non-reduced scCNA data.

However, a better performance was achieved for all dataset sizes when clustering was applied following feature selection and dimensionality reduction. This confirms that this data's high-dimensional and noisy nature negatively affects clustering results. In this scenario, PCA preprocessing was more effective when dealing with smaller datasets, while UMAP worked better with larger ones. It is generally believed that clustering following UMAP embeddings should be avoided since UMAP affects the global data structure while maintaining the local relationships between data points [145]. UMAP can also create false tears in clusters, resulting in excessively fined grained clustering. Despite these concerns, there are still valid reasons to use UMAP as a preprocessing step before clustering. Specifically, UMAP is particularly effective in uncovering the underlying signals from data with a vast number of dimensions, most of which are noisy or redundant. When this is the case, UMAP preprocessing may be therefore beneficial, provided that a manual inspection of the results is performed [146].

Indeed, at least in this experiment, on average, the best performance was obtained when applying UMAP preprocessing, particularly when combined with density-based clustering approaches, which suggests that UMAP preprocessing may be helpful to reduce scCNA data dimensionality before clustering.

In general, it is worth noting that the clustering methods that provided, on average, the most accurate results are those that do not require seeded with the cluster number. This may be a consequence of the automatic selection of the K, determined by maximizing the Silhouette score. This allows to conclude that, when dealing with large-scale and high-dimensional data, where the number of clusters is unknown, clustering methods that can infer the number of clusters from the data are always the best choice.



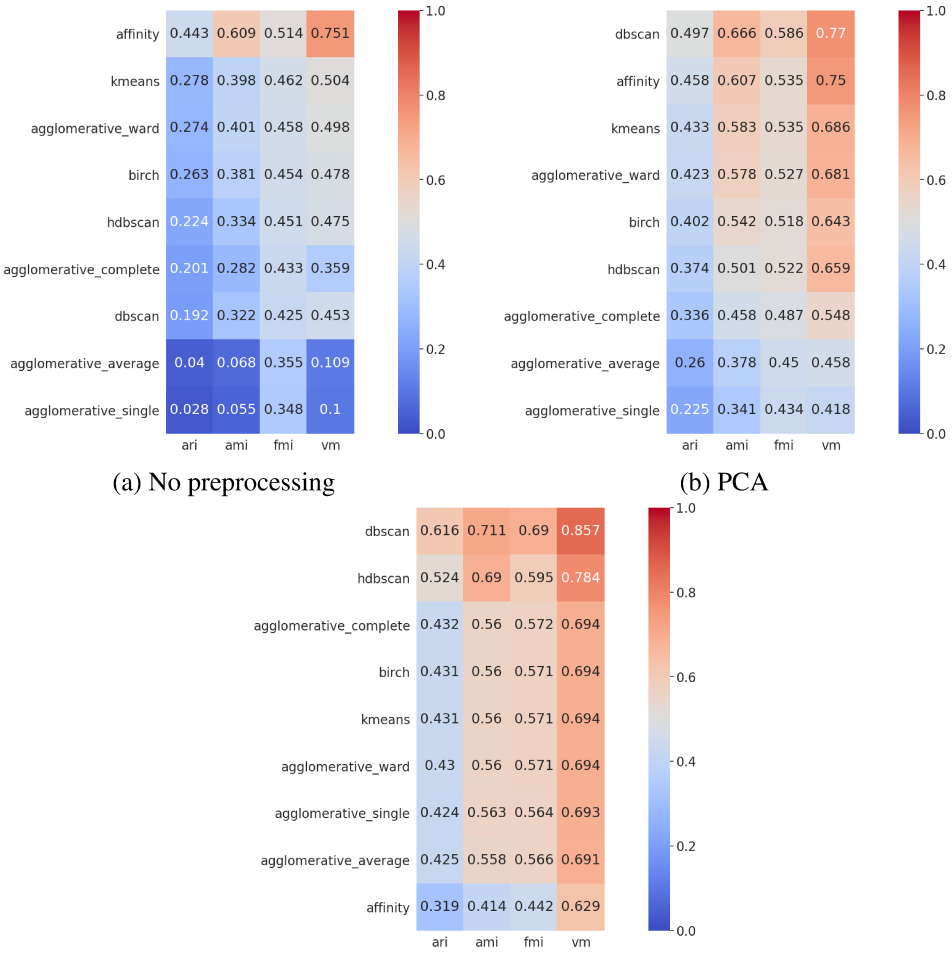
(a) No preprocessing

(b) PCA

(c) UMAP

Fig. 4.2 Clustering algorithm accuracy on 100 cells dataset: mean external validation indices scores. Clustering algorithms are sorted according to the average ranking computed on all indices.

Moreover, to obtain an indication of the average accuracy of each algorithm in the three preprocessing scenarios, the indices were rescaled to the interval $[0,1]$, and a mean accuracy score was computed across the dataset sizes. Table 4.2 shows that UMAP should be preferred over PCA, especially when used before running DBSCAN or HDBSCAN. On the other side, to exploit the full resolution of the data, AP is the most accurate algorithm.



(a) No preprocessing

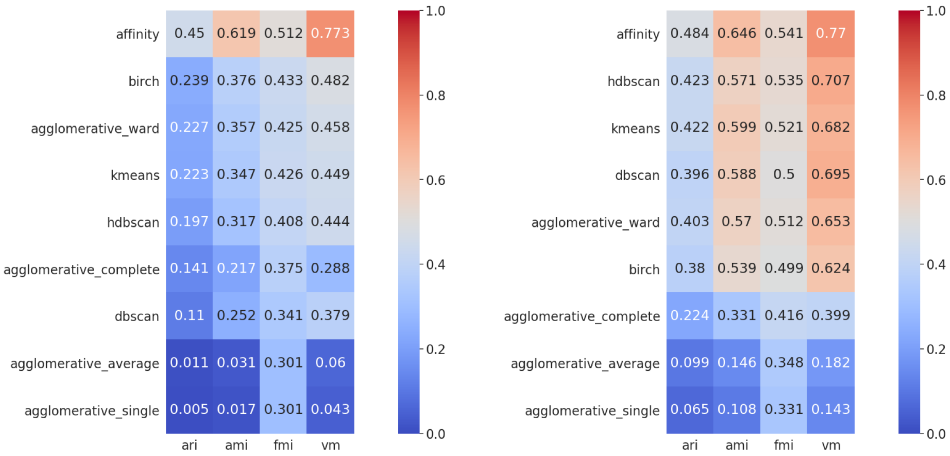
(b) PCA

(c) UMAP

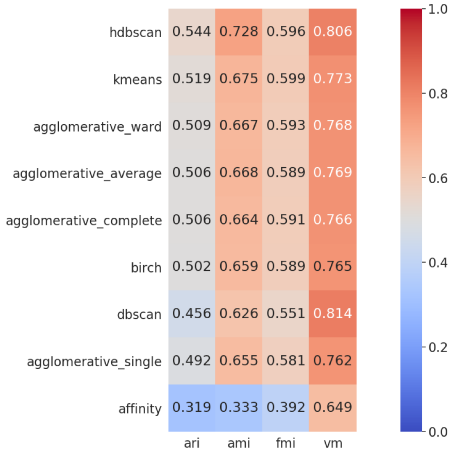
Fig. 4.3 Clustering algorithm accuracy on 200 cells dataset: mean external validation indices scores. Clustering algorithms are sorted according to the average ranking computed on all indices.

4.3.2 Test case: SW480 cells

The algorithms achieving the best performance in the experiment with 400 cells were tested on SW480 cell data. As a general preprocessing step, the cells characterized by a high MAD ($> 90^{\text{th}}$ percentile) were filtered out. The MAD is the median absolute deviation of all pair-wise differences in read counts between neighboring bins and reflects the bin count dispersion due to technical noise. After that, Affinity Propagation (AP) clustering was applied to the non-reduced dataset and HBSCAN to UMAP-preprocessed data. In order to determine the model with better separation



(a) No preprocessing (b) PCA



(c) UMAP

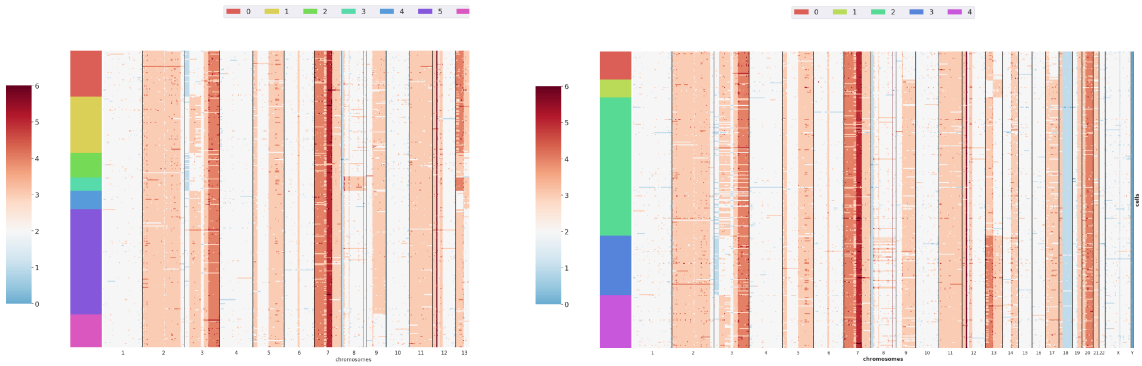
Fig. 4.4 Clustering algorithm accuracy on 400 cells dataset: mean external validation indices scores. Clustering algorithms are sorted according to the average ranking computed on all indices.

between the clusters, the Davies-Bouldin score [147] was computed (lower values signifies better cluster separation).

Figures 4.5a and 4.6a show AP results. Clusters composed of less than 10 items were excluded, and only the 7 major clusters were kept for further analysis. The scatter-plot (Figure 4.6a) shows that the clusters were well separated, except for a few cells which have been misclassified. The heatmap shows that clusters were internally cohesive, and each of them contained CNA profiles that were distinguishable from those of the other clusters. The Davies-Bouldin score had a value of 9.933.

	No preproc.	PCA	UMAP
dbscan	0.401	0.715	0.783
kmeans	0.467	0.627	0.556
hdbscan	0.435	0.606	0.698
affinity	0.723	0.658	0.390
aggl. ward	0.465	0.602	0.557
birch	0.462	0.572	0.546
aggl. complete	0.351	0.460	0.550
aggl. average	0.164	0.364	0.547
aggl. single	0.141	0.298	0.538

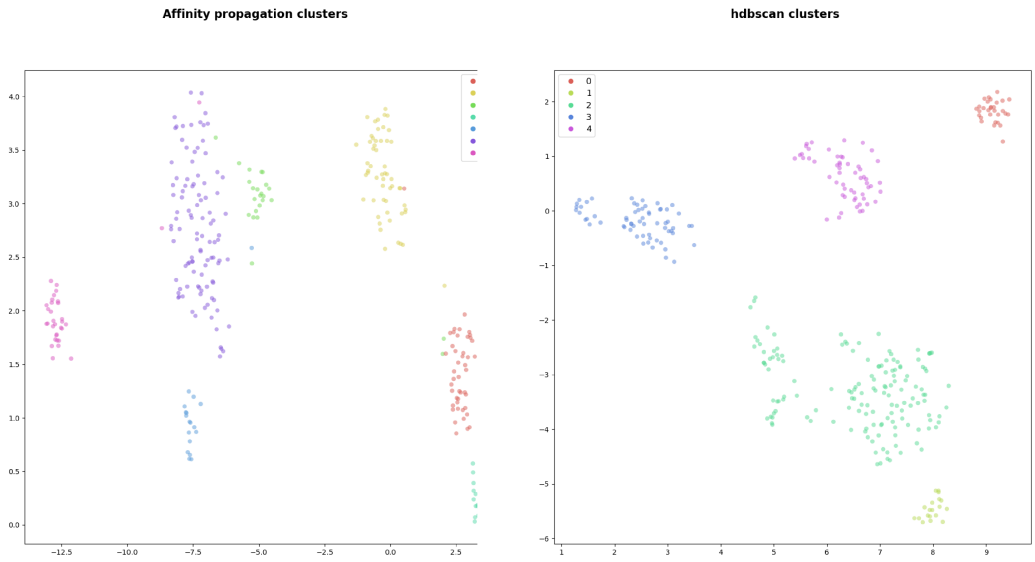
Table 4.2 Clustering algorithm evaluation: overall mean accuracy scores.



(a) AP (b) UMAP+HDBSCAN

Fig. 4.5 SW480 clusters. We applied AP on the non-reduced dataset (4.5a) and HBSCAN on the UMAP-reduced one (4.5b). The colored labels on the left-side of the heatmaps indicate the cluster which each cell was assigned to.

Figures 4.5b and 4.6b show HDBSCAN results. The cells marked as "noise" by the algorithm were excluded. Additionally, the HDBSCAN library implements the GLOSH outlier detection algorithm, which can detect outliers that may be noticeably different from points in its local region (for example, points not on a local submanifold) but that are not necessarily outliers globally. So we took advantage of this feature to filter out also the cells with a high outlier-score ($> 90^{\text{th}}$ percentile). In the end, we obtained 5 clusters. The scatter-plot (Figure 4.6b) shows that, in this case, all cells were assigned to the most appropriate cluster. The heatmap (Figure 4.5b) shows quite consistent clusters, even if clusters 2 and 3 could have been split into two subclusters. The Davies-Bouldin score had a value of 10.950.



(a) AP (b) UMAP+HDBSCAN

Fig. 4.6 2D representation of clustering results (without outliers). The 2D representation of the dataset shows that AP (4.6a) assigned a few cells to the wrong cluster, while HDBSCAN (4.6b) failed in splitting cluster 2 and 3.

Remarkably, the clusters returned by the two methods applied once on the raw dataset and once after an aggressive dimensionality reduction performed with UMAP are very similar. This means that UMAP may be used as a preprocessing step for clustering, as long as some manual validation of the results is performed. AP produced more clusters than HDBSCAN because it separated the cells, which the latter algorithm put into clusters 2 and 3, as reflected by the Davies-Bouldin score. On the other side, HDBSCAN is cluster shape independent and is resilient to noise and outliers.

4.4 Conclusions

The task of evaluating clustering algorithms' performance on scCNA data has shown to be challenging and insidious, mainly because ground truth information about cell phylogeny is not available for public scDNAseq datasets.

Here, a synthetic dataset of single-cell CNA profiles with a known underlying phylogeny has been used to perform the first formal and systematic evaluation of

clustering algorithms onto single-cell CNA data, raising some data science issues. The performance of nine well-known clustering algorithms was compared, highlighting the pros and cons of the methods in predicting the structure of the real cell phylogeny. Three different dataset sizes were considered, and both situations in which data are reduced to a lower-dimensional space (PCA/UMAP) and when they are not were tested. The computation time, algorithm stability (APN), and algorithm accuracy (ARI, AMI, FMI, VM) were computed for each algorithm run. They all produced highly stable results, while density-based algorithms are those in which computation time increases more rapidly by increasing the dataset size. As for the accuracy, the algorithms were ranked based on the average of the four indices. The algorithms that do not require to be seeded with the cluster number outperformed the others. Specifically, Affinity Propagation won when no dimensionality reduction was performed, while density-based algorithms had outstanding results on top of PCA and UMAP results (DBSCAN for 100 and 200 cells dataset, HDBSCAN for 400 cells dataset).

Affinity Propagation and HDBSCAN were tested on a real scCNA dataset. AP was applied on the non-reduced dataset, while HDBSCAN was performed following UMAP preprocessing. They both extracted cohesive and well-separated clusters. Moreover, the clusters identified by the two algorithms were similar, suggesting that UMAP may be effectively exploited to perform dimensionality reduction. AP outperformed HDBSCAN in separating the items of two subgroups, which may indicate that retaining the complete set of features may increase the resolution in subclones identification.

The main limitation of the present work is that the algorithm benchmarking was performed on synthetic data due to the lack of an available biological ground truth; for this reason, an ad-hoc experiment should be designed to produce real data and extend our analysis.

To conclude, this work has presented a framework to study clustering algorithms' performance on scCNA data, which can be easily replicated to perform similar studies.

Chapter 5

PhyliCS: A Python Library To Explore scCNA Data And Quantify Spatial Tumor Heterogeneity

Tumors are composed of a number of cancer cell subpopulations (subclones), characterized by a distinguishable set of mutations. This phenomenon, known as intra-tumor heterogeneity (ITH), may be studied using Copy Number Aberrations (CNAs). Nowadays, ITH can be assessed at the highest possible resolution using single-cell DNA (scDNA) sequencing technology. Additionally, single-cell CNA (scCNA) profiles from multiple samples of the same tumor can, in principle, be exploited to study the spatial distribution of subclones within a tumor mass. However, since the technology required to generate large scDNA sequencing datasets is relatively recent, dedicated analytical approaches are still lacking.

This chapter presents PhyliCS, the first tool which exploits scCNA data from multiple samples from the same tumor to estimate whether the different clones of a tumor are well mixed or spatially separated. Starting from the CNA data produced with third-party instruments, it computes a score, the SHscore (Spatial Heterogeneity score), to distinguish spatially intermixed cell populations from spatially segregated ones. Additionally, it provides functionalities to facilitate scDNA analysis, such as feature selection and dimensionality reduction methods, visualization tools, and a flexible clustering module.

PhyliCS represents a valuable instrument to explore the extent of spatial heterogeneity in multi-regional tumor sampling, exploiting the potential of scCNA data.

5.1 Scientific background

Tumors are caused by the accumulation of somatic mutations. The set of mutations accumulated by the founder cell of a tumor is defined as clonal and inherited by its entire progeny. The mutations arising in an already existing tumor are passed on only to sub-populations of cells and are defined as subclonal [36, 37]. As a result, cancer cells are characterized by an intrinsic genetic diversity, known as intra-tumor heterogeneity (ITH) [38].

ITH is a major topic of interest for the cancer research community since it has been recognized as one of the principal responsible for tumor relapse and treatment failure [39, 38, 40–42]. The most common way to assess ITH is to use deconvolution techniques on bulk DNA sequencing data [49, 50]. Such techniques are generally based on machine learning models, used to cluster the mutations into subclones based on their prevalence and exploit such clusters to infer the tumor phylogenetic structure [53–61]. Some studies have proposed methods to evaluate ITH based on gene expression [73–75] or protein-protein interactions [76].

Several studies have shown that using multiple samples taken from distinct regions of the same lesion improves the ability to infer the subclonal structure of tumors [44–47, 38–40, 47, 48, 148] and assess ITH. For example, a study conducted by Jamal-Hanjani et al. [149], sampling 327 regions from 100 early-stage non-small-cell lung cancers, revealed that 30% of the somatic mutations were subclonal and stated that if fewer regions had been sampled, many of those mutations would have misinterpreted as clonal.

In this context, emerging single-cell DNA sequencing (scDNA-seq) technologies offer an extraordinary opportunity to tackle such issues, as they allow to study tumor heterogeneity with unprecedented resolution. In particular, single-cell low-coverage whole-genome sequencing is suited for detecting chromosomal aberrations, which can be exploited to reconstruct cell population subclonal structure [62].

However, the existing methods for single-cell CNA (scCNA) analysis are still limited. Many of them [64–71] only identify the total copy-number, which indicates the sum of the number of copies at each locus, by analyzing differences between the observed and expected number of sequences aligned to a locus, or the read-depth ratio. A few of them also infer the tumor phylogeny using the CNAs they computed [72].

However, an instrument capable of exploiting both the granularity of single-cell DNA data and multi-sample analysis to quantify ITH still does not exist.

Therefore, PhyliCS is presented: it is a flexible Python library that explores CNA calls obtained with third-party tools and exploits them to compute a new metric, the Spatial-Heterogeneity score (SHscore). This score is helpful to evaluate the spatial heterogeneity of tumors when multiple regional samplings are available, quantifying how much cells from different samples from the same patient have diverged in their CN landscapes. This evaluation allows both to rank different tumors based on their heterogeneity and to identify the most divergent spatial samples of a given tumor. Additionally, it may help to explore different tumors without a huge number of sequenced cells and/or regional samplings to select only the most heterogeneous ones for further analyses.

Moreover, PhyliCS provides easy access to several clustering methods for both single and multiple samples to users, making it easy to compare results and tailor each analysis to each specific experiment. Its potential is shown by running it on 300 simulated datasets to validate the SHscore on some selected ideal scenarios where it compares sets of cells with known relationships. After that, the correlation between the proposed SHscore and the evolutionary distance between the cells of the samples in analysis is demonstrated through a more extensive simulation experiment. Lastly, PhyliCS has been tested on three publicly available scDNA datasets: one with multiple spatial samplings from a breast tumor, another comprised of a primary lung tumor and its derived metastases, and a third with a cell line and two clonal expansions of two single cells. The last part of this chapter describes the results of this analysis, using the SHscore to describe how the CN profiles differ when considering the fine-grained single-cell level in the bigger context of multiple sampling.

5.2 Implementation

This section describes the main modules of PhyliCS and presents the mathematical details of the SHscore and its interpretation.

5.2.1 PhyliCS

PhyliCS is a comprehensive toolkit integrating scCNA calls analysis procedures into a single and modular Python package.

As Figure 5.1 shows, PhyliCS takes as input the scCNA calls produced by one of the existing scCNA callers [64–72] and allows the users to perform:

- data preprocessing (feature selection, PCA, UMAP, data filtering),
- data visualization (UMAP-based scatterplots, heatmaps),
- data clustering (Affinity Propagation [130], Birch [132], DBSCAN [133], HDBSCAN [134], Hierarchical Agglomerative [131], KMeans [135], OPTICS [150], Spectral [151]),
- clustering algorithms evaluation (Silhouette Coefficient, Davies-Bouldin Index, Calinski-Harabasz Index, Adjusted Rand Index, V-Measure, Fowlkes-Mallows Score, Mutual Information),
- multi-sample clustering, visualization and spatial intra-tumor heterogeneity estimation (SHscore).

PhyliCS multi-sample analysis module works on the aggregation of input sample data and produces two main results: a graphical representation and a numerical quantification of spatial intra-tumor heterogeneity, the SHscore. Specifically, it generates an aggregated heatmap with a dendrogram computed performing hierarchical clustering of the cells. Different colored labels identify heatmap rows representing the cells from the different samples. In this way, it is possible to assess whether the clustering algorithm segregated cells originating from different samples into different branches of the dendrogram or if generated mixed clusters: the former case would indicate that, despite originating from the same tumor, the genomic make-up of the cells belonging to different samples is different (spatial intra-tumor heterogeneity);

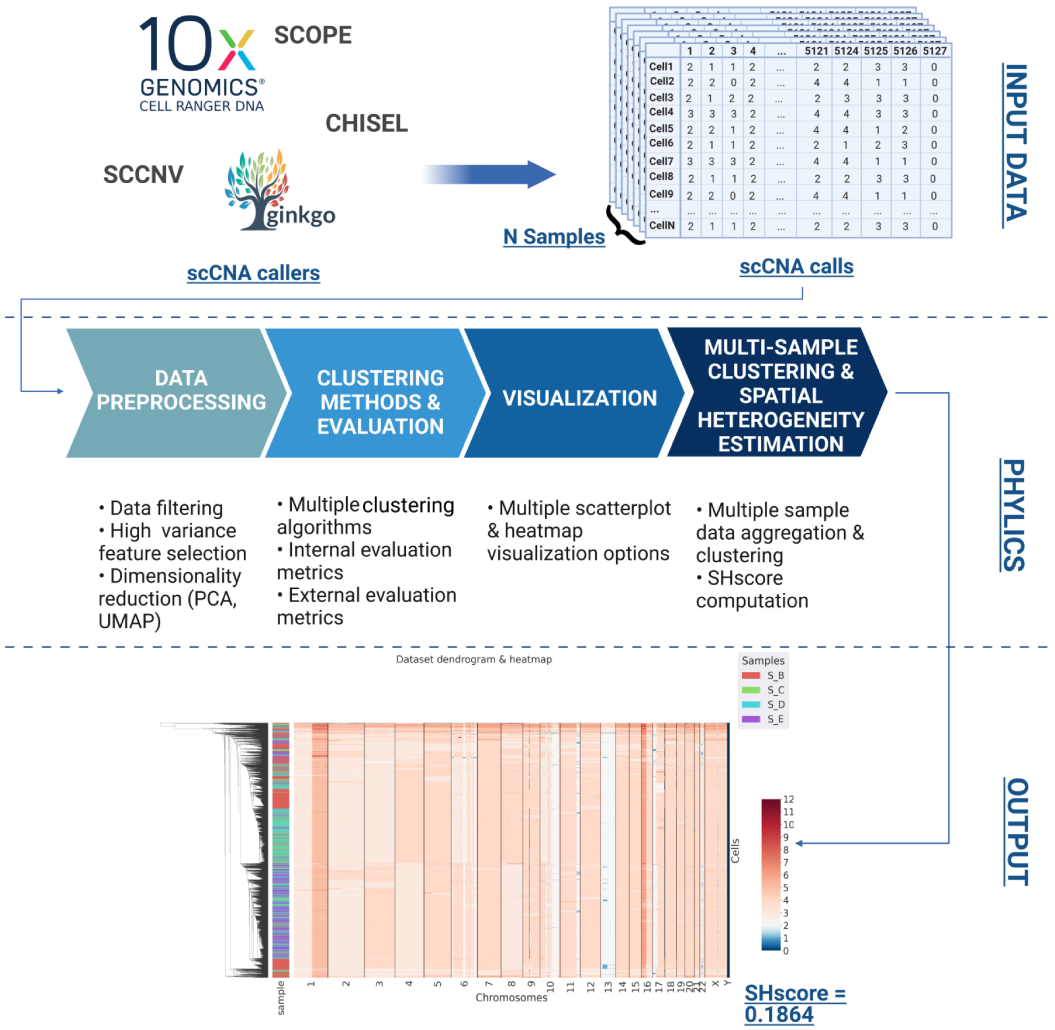


Fig. 5.1 PhyliCS logical schema. PhyliCS allows to perform downstream analysis on the scCNA profiles computed with scCNA third party callers. Specifically, it accepts tabular data and allows to perform data filtering, feature selection, dimensionality reduction, to prepare the data before executing one of the multiple available clustering algorithms. It also allows to perform clustering result quality evaluation by means of both internal and external evaluation metrics. But, most importantly, it provides the possibility to aggregrate scCNA data from multiple samples, to jointly cluster and visualize them, estimating their spatial ITH through the SHscore.

the latter case, instead, would denote that different samples are populated by cells with a similar genomic variance.

PhyliCS implementation is based on a dedicated class, named *CnvData*, which is a modular data structure storing all data annotations (e.g., cell ploidy, cell MAD, etc.)

and the results of each analytical step (e.g., PCA, clustering results, etc.) without affecting the data matrix. On the one side, this implementation choice simplifies and speeds up computation; on the other side, it allows experienced developers to extend the framework and add new functionalities with a low programming effort.

PhyliCS does not represent an alternative to the existing scCNA tools developed for identifying scCNA events [64–71] or tools designed for the phylogenetic analysis [72]. Indeed, PhyliCS offers an API to work on scCNA data, leveraging different third-party tools' outputs and implementing a method to characterize spatial ITH.

5.2.2 Spatial Heterogeity Score

The Spatial-Heterogeneity score (SHscore) is a relative measure of how much the genomic make-up of different samples taken from the same patient diverges with respect to the internal variance of each sample.

Definition The principles underlying the SHscore are inspired by those of the Silhouette score, an index used in classical Data Science, to estimate the quality clustering results [152]. Cells can be thought of as data points described by their CNA profile and the samples as the cluster they belong to. It is possible to compute for each cell, p, the average distance from all other cells belonging to its cluster, $a(p)$, and then compare it to the average distance from the cells belonging to the "nearest", or most similar, cluster, $b(p)$. Figure 5.2 shows a conceptual schema of a tumor divided into two subsamples: green arrows represent the pairwise distance between a given cell, p, and all cells of its sample; the orange ones, the distance between the same cell and cells of the nearest sample. The average computed on these distances are $a(p)$ and $b(p)$.

These distances are the same used to compute the Silhouette score, so its implementation has been adapted to the purposes of this work.

For each cell p and sample S_p, such that $p \in S_p$, let $a(p)$ (*Equation 5.1*) be the average pairwise-distance between p and the other cells belonging to its sample and $b(p)$ (*Equation 5.2*) be the minimum average pairwise-distance between p and other sample cells. Now, $sh(p)$ (*Equation 5.3*) can be computed: it measures the difference between the average pairwise-distance between p and the cells of the sample, nearest

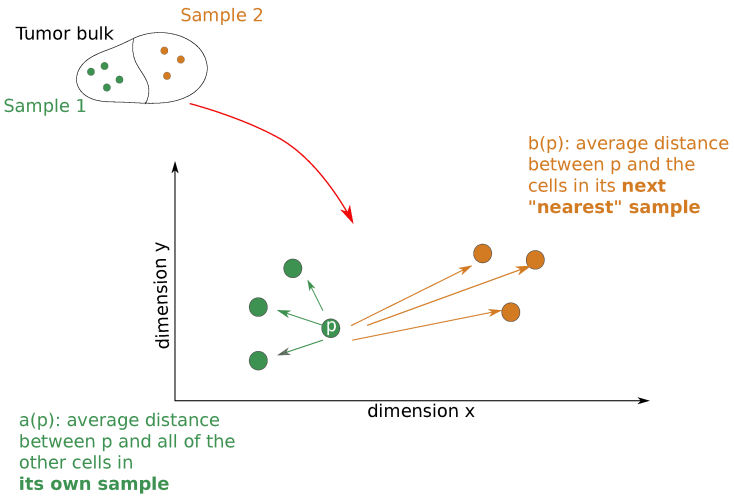


Fig. 5.2 Intra and inter-sample pairwise distance. Given the cell p, $a(p)$, is the average pairwise distance between p and all other cells from its own sample, while $b(p)$ is the average pairwise distance between p and the cells from the "nearest" sample.

to the one it belongs to, and the average pairwise-distance between p and the cells of its own sample.

$$a(p) = \frac{1}{\mid S_p \mid -1} \sum_{p \in S_p, q \neq p} d(p,q) \tag{5.1}$$

$$b(p) = min_{k \neq p} \frac{1}{\mid S_k \mid} \sum_{q \in S_k} d(p,q) \tag{5.2}$$

$$sh(p) = \frac{b(p) - a(p)}{max\{a(p), b(p)\}} \tag{5.3}$$

Dividing it by $max\{a(p), b(p)\}$ makes $sh(p)$ a relative difference.

In order to mitigate the negative impact of the high dimensionality of scCNA data, we adopted *L1*, or *Manhattan*, norm to compute pairwise distances. In fact, it has been demonstrated that, for dimensionalities of 20 or higher, *LK* norms, with $K \leq 1$, better discriminate [153, 154] between the nearest and the furthest neighbors compared to higher level norms (e.g. *L2*, or *Euclidean* norm).

From *Equation 5.3* it is clear that $-1 \leq sh(p) \leq +1$.

For $sh(i)$ to be close to 1 we require $a(p) << b(p)$. As $a(p)$ measures how much the genomic profile of p is dissimilar to the average profile of its sample, a small

value means a high level of similarity. Furthermore, a large $b(p)$ indicates that p CNA profile is highly different from the average profile of the most similar among the samples in the analysis. Thus, a $sh(p)$ close to 1 means that p CNA profile matches the average genomic profile of the sample it belongs to. If $sh(p)$ is close to -1, then by the same logic, it is possible to state that p CNA profile is more similar to the genomic profile of the neighboring sample than to the genomic profile of the other cells of its sample. An $sh(p)$ close to 0 means that the CNA profile is on the border of two natural clusters, so p may belong to both of them.

Mathematically, the SHscore, $SHscore(S_1, S_2, ..., S_n)$, for the set of samples $S_1, S_2, ..., S_n$, is a measure of how well-separated the samples are and is defined as the mean $sh(p)$ over all cells in the entire dataset, $D = [S_1 \cup S_2 \cup ... \cup S_n]$ (*Equation 5.4*).

$$SHscore(S_1, S_2, ..., S_n) = \frac{\sum_{p, p \in D} sh(p)}{\mid D \mid}. \tag{5.4}$$

From *Equation 5.4*, it is clear that also the SHscore may assume values in the interval $[-1, 1]$ and its interpretation may be derived from the interpretation of single-cell scores. Specifically, a SHscore close to 1 indicates that many cells in the various samples are characterized by a $sh(p)$ close to 1, denoting that samples are internally homogeneous and segregated with respect to the others. Similarly, a SHscore close to -1 indicates that many cells in the dataset look more similar to the cells of another sample than those of their sample; this could denote problems with the sequencing quality or data pre-processing. Finally, a SHscore close to 0 implies that many cells may indistinctly belong to their sample or to another, which may indicate two scenarios: the samples are internally homogeneous but very similar among each other; thus, they share the same subclonal structure and cells may belong to one or another; or that the samples are internally heterogeneous so that the CN profiles of their cells cannot be assigned to any one of them.

Application scenario Let us suppose that three single-cell data-sets, s_1, s_2, s_3, originated from three different regions of the same tumor, have been provided as input samples to PhyliCS. The SHscore evaluation phase will proceed as follows:

1. The cells are assigned to three predefined clusters, S_1, S_2, S_3, in the following way: $\{p : p \in s_i\} \Rightarrow p \in S_i$, where $i \in [1, 2, 3]$. The SHscore is computed as $hs_{1,2,3} = SHscore(S_1, S_2, S_3)$

2. The cells from s_1 and s_2 are combined in a single cluster, S_{12}, and those from s_3 are assigned to a separate cluster, S_3. The SHscore is computed again as $hs_{12,3} = SHscore(S_{12}, S_3)$.

3. The cells from s_1 and s_3 are combined in a single cluster, S_{13}, and those from s_2 are assigned to a separate cluster, S_2. The SHscore is computed again as $sh_{13,2} = SHscore(S_{13}, S_2)$

4. The cells from s_2 and s_3 are combined in a single cluster, S_{23}, and those from s_1 are assigned to a separate cluster, S_1. The SHscore is computed again as $sh_{23,1} = SHscore(S_{23}, S_1)$.

Let us suppose, now, that $hs_{23,1}$ is the maximum computed score. Specifically, we suppose that:

$$sh_{23,1} > sh_{1,2,3}. \tag{5.5}$$

This means that samples S_2 and S_3 are similar and, in some measure, different from sample S_1 and that considering their cells together resulted in a better clustering.

To conclude, the SHscore represents a way to quantify the genomic distance numerically, in terms of CNAs, between different samples of the same tumor and to investigate spatial intra-tumor heterogeneity.

5.3 Results and Discussion

Here, the experiments conducted to study the SHscore behavior in different contexts are introduced. Additionally, the procedures executed to generate the simulated datasets are described.

In detail, the SHscore has been used on 200 simulated datasets representing some ideal scenarios (spatial segregation, spatial intermixing, early metastasis spreading, and late metastasis spreading) to check if it correctly reflects the heterogeneity in the clonal structure of multiple samples. After that, the score has been tested on a set of 100 simulations to analyze its behavior when the mean CNA size and the mean number of copies gained varies in a controlled way. Then, a more extensive simulation was conducted to verify the correlation between the SHscore and the divergence accumulated during the evolution of the samples. Finally, the SHScore

has been tested on 3 publicly available scCNA datasets to study its behavior in some real-world scenarios.

5.3.1 Experiment 1: SHscore on synthetic data

Data generation

A simulation study was conducted to analyze the SHscore behavior under four different scenarios (spatial subclone segregation, spatial subclone intermixing, early and late metastasis spreading) and to study if and how it correlates with some features of the CN profiles of cells (CNV region size, CN level).

To this purpose, the model presented by Fan et al. [155] to generate a phylogenetic tree starting from a reference genome, using a generalization of the Beta-Splitting model [126], was extended. At the end of the simulation process, the leaves of the generated tree represent the cells sampled from the patient, while the internal nodes represent intermediate CN states, which do not exist anymore.

Spatial segregation To simulate the extreme case in which subclones segregate in isolated niches very early during tumor evolution, the progeny of the first 5 cells (Figure 5.3a) generated by the simulator was tracked. The trees grew until they contained 2500 leaves. At that point, the groups of phylogenetically separated cells were distinguishable and could be considered as subsamples, each containing a distinct subclone (Figure 5.3b). So, in the end, each dataset was divided into 5 subsamples corresponding to the 5 groups of cells deriving from the first 5 generated cells. From now on, this scenario is referred to as *hom-scenario*.

Spatial intermixing Another experiment simulated the scenario in which the tumor cells subpopulations are spatially well-mixed so that a regional subsampling would produce very similar samples. This was done by shuffling the leaves of the previously generated trees and randomly assigning them to 5 subsamples (Figure 5.3c). From now on, this scenario is referred to as *het-scenario*.

Metastasis spreading Another and different case of spatial segregation was simulated: the scenario in which a cell seeds a metastasis, initiating a completely isolated

clonal expansion. To that purpose, new phylogenetic trees were generated: when the trees had generated 1/4 or 3/4 of the final number of cells, one cell was randomly selected, and another tree was seeded to model early or late metastatic spreading during the primary tumor evolution, respectively. The tree generation proceeded in parallel until all of them contained 500 leaves (Figure 5.4). From now on, these scenarios are referred to, respectively, as *early-met-scenario* and *late-met-scenario*.

For each of the four scenarios described so far, 50 synthetic datasets were generated for a total of 200 simulations.

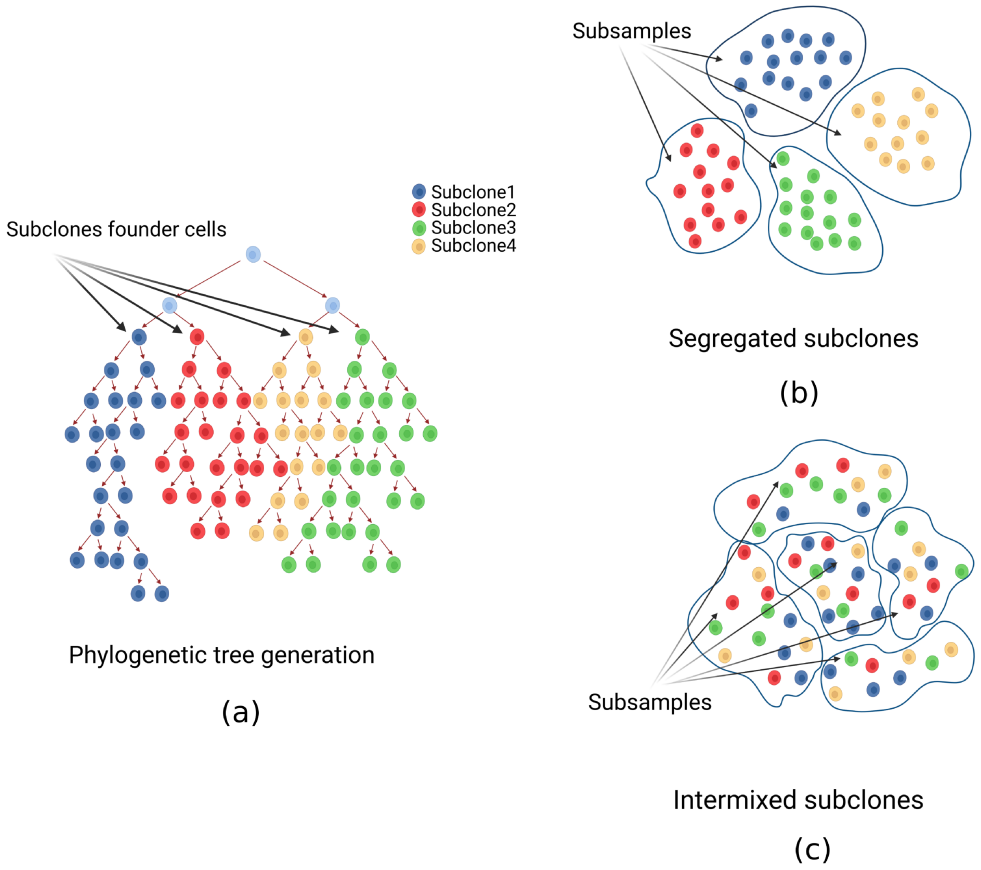


Fig. 5.3 Spatial subclonal segregation and intermixing simulation. 50 phylogenetic trees (3a) made of 2500 cells were generated. For each tree, two scenarios were simulatated: (I) early segregation of subclones (*hom*) by tracking the progeny of the first five generated cells and assigned the leaves to five distinct subsamples, corresponding to the five subclones (3b); (II) spatial intermixing of subclones (*het*) by shuffling the leaves and assigning them randomly to five subsamples (3c).

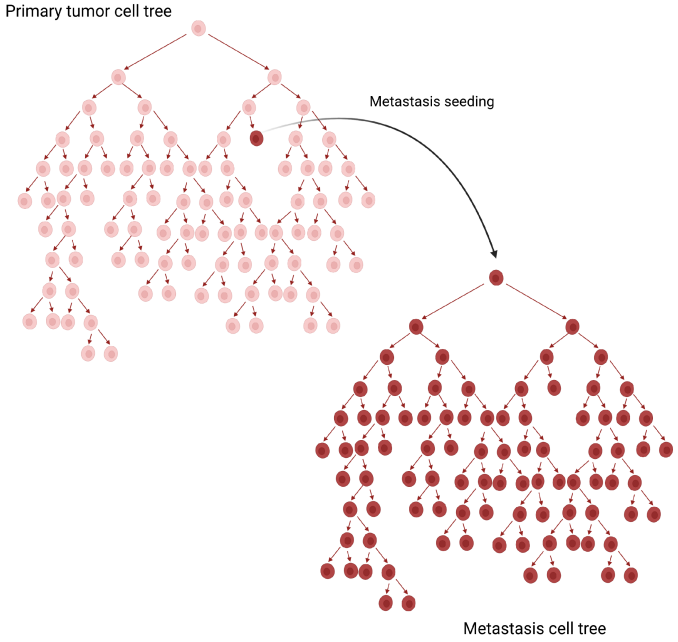


Fig. 5.4 Metastasis seeding and expansion simulation. 100 pairs of primary-metastasis samples were generated (50 *early* metastasising, 50 *late* metastasising). Each pair was obtained by seeding the primary tumor tree and successively initiating a new tree with a cell randomly selected when the primary tree had generated 1/3 (early) or 3/4 (late) of the final number of desired cells. The simulation was stopped when both trees had generated 500 leaves.

Simulations with varying parameters 100 datasets were simulated with varying parameters to generate CN profiles characterized by different structural features and check if and how those features correlate to the SHscore. Precisely, two parameters were varied: the expected CNA size (θ), used by the simulator to sample from an exponential distribution, and the reciprocal of the expected number of gained copies (p), used to sample from a geometrical distribution. In details, for each simulation, θ was chosen by randomly sampling from a uniform distribution defined in the interval $[500, 5000000]$, while p was sampled from a uniform distribution defined in the interval $[0.1, 0.9]$. Each simulated tree had 1000 leaves and was split into two subtrees representing a tumor subsample. From now on, this scenario is referred to as *var-scenario*.

SHscore statistics

SHscore was computed on the synthetic datasets, built to represent the previously described heterogeneity scenarios, to evaluate its ability to capture their differences.

Spatial heterogeneity at the same disease site First, the SHscore was computed on the 100 sets of samples simulating the regional subsampling from the same disease site. (Figures 5.5a and5.5b). Figure 5.5a shows the SHscores computed on the *hom-scenario* (spatial segregation) and the *het-scenario* (intermixing). The scores, in the two scenarios, are different (unpaired wilcoxon p-value 3.5×10^{-18}): in the *het-scenario* values fall into a very small interval (min: -0.020, max: -0.004, median: -0.010, IQR: 0.004); the *hom-scenario*, instead, produced scores ranging on a broader interval (min: 0.043, max: 0.295, median: 0.151, IQR: 0.064), reflecting a higher heterogeneity between the simulated samples with different "clones" (the progenies of the first five cells) evenly distributed among them.

The results obtained by this experiment demonstrated that the proposed score is able to discriminate between the two described scenarios.

Spatial heterogeneity at different disease sites Figure 5.5b shows the results for the two metastatic scenarios: here too the difference is significant (Mann–Whitney U p-value 0.0029), albeit less pronounced, underlying how different seeding histories can result in different SHscores; even if with the parameters chosen for the simu-

lations the difference is small and the intra-scenario variability between different simulation is high (*early-met*: min: 0.103, max: 0.461, median: 0.267, IQR: 0.124; *late-met*: min: 0.195, max: 0.547, median: 0.320, IQR: 0.084).

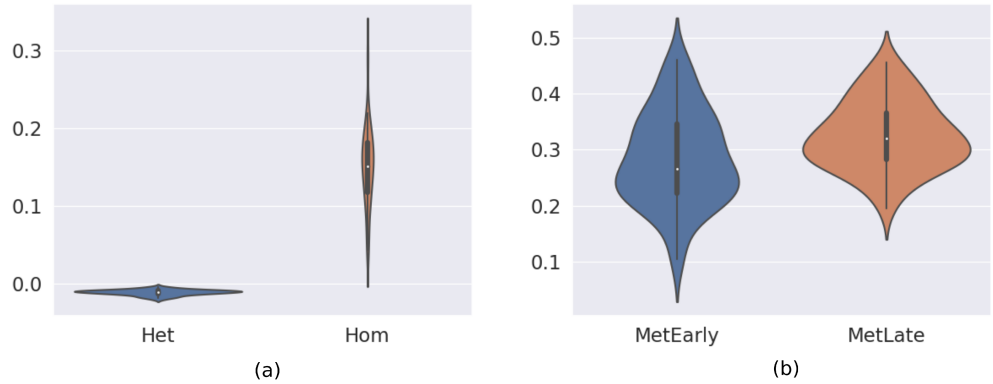


Fig. 5.5 SHscore distribution. The SHscore was computed on 100 synthetic datasets simulating regional subsampling (5a) (Mann–Whitney U test p-value 3.5×10^{-18}). Het-scenario = min: -0.020, max: -0.004, median: -0.010, IQR: 0.004. Hom-scenario = min: 0.043, max: 0.295, median: 0.151, IQR: 0.064. We also computed the SHscore on 100 synthetic dataset simulating metastasis spreading (5b)(Mann–Whitney U p-value 0.0029). EarlyMet scenario = min: 0.103, max: 0.461, median: 0.267, IQR: 0.124; LateMet scenario = min: 0.195, max: 0.547, median: 0.320, IQR: 0.084.

SHscore indipendence from CNA size and gained copy number In order to study if the SHscore correlates with the mean CNA size and the mean number of gained copies, the SHscore for each pair of samples generated in the *var-scenario* was computed. The Pearson correlation coefficient between the SHscores and the parameters θ (mean CNA size) and p (reciprocal of mean number of gained copies) was computed for each simulation. The results (θ: Pearson correlation coefficient = -0.101, pvalue =0.319; p: Pearson correlation coefficient = -0.109, pvalue = 0.282) indicated that there were no significant correlations, suggesting that SHscore is robust with respect to different rates of CN accumulation and to the size of events.

5.3.2 Experiment 2: SHscore and evolutionary distance

The heterogeneity quantified by the SHscore reflects the evolutionary distance between the cells of the samples analyzed. Another simulation experiment was designed

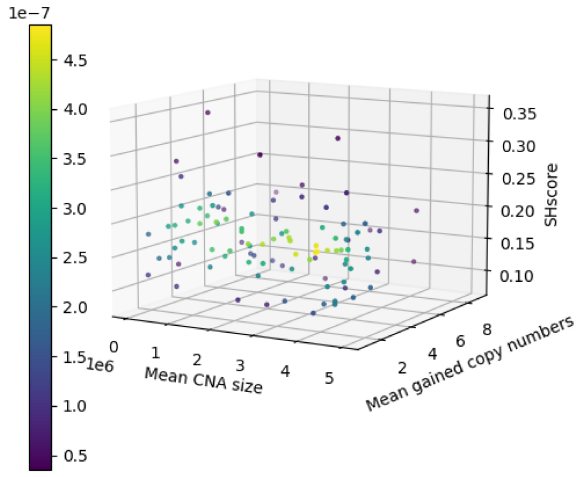


Fig. 5.6 SHscore independence from mean CNA size and mean gained copies. The SHscore was tested on multiple simulated sample pairs, characterized by a different and known mean CNA size θ and mean number of gained copies p. The results show that the SHscore is uncorrelated to those features with a Pearson correlation coefficient c = -0.101 (pvalue = p=0.319), for the mean CNA size, and c = -0.109 (pvalue = 0.282), for the mean number of gained copies (notice that plots $1/p$)

to verify the existence of a correlation between SHscore and the distance between the copy-number states, which originated the mutational profile of the samples. Such CN states may be thought of as the most recent common ancestor (MRCA) of the existing CN profiles.

Data generation

100Kcells and 10Kcells. In order to generate a deep evolutionary history and, consequently, a more heterogeneous dataset, a cell-division tree with 100K final leaves was simulated. The subtrees rooted in the first 200 generated cells were tracked, simulating the complete spatial segregation of the subclones originating from those cells (see Spatial heterogeneity at the same disease site). The cardinality of the generated datasets was relatively homogeneous (mean cell number = 500 cells) with some exceptions (min cell number = 91, max cell number = 3112). In order to have a balanced dataset, only the subtrees with a cardinality between the 1^{st} and the 3^{rd} quartile (208.75 and 746.50 leaves, respectively) were retained. For each subtree, the leaves were extracted, and the CNA matrix was generated; additionally,

the position of their roots within the parental tree was tracked. From now on, we refer to this scenario as the *100Kcells* experiment.

The same procedure was executed to generate trees with 10K leaves, tracking subtrees for the first 20 generated cells. Also, in this case, only the datasets with a cardinality between the 1^{st} and the 3^{rd} quartile (318 and 623.75 leaves, respectively) were kept. From now on, this scenario is referred to as the *10Kcells* experiment.

SHscore and MRCA distance correlation

In order to investigate the correlation between the SHscore and the distance between the MRCAs of the sample cells, the dataset generated in the *100Kcells* experiment was used. First, the SHscores were computed for the 4950 possible pairs of samples. After that, 1000 pairs were randomly selected, and the distance between their MRCAs, represented by the number of edges connecting the single cells that originated the two subtrees, was computed. The random selection was representative of the whole set of pairs since they were equally distributed (Kruskal-Wallis pvalue = 0.941, Figure 5.7).

Finally, it was possible to demonstrate that the two quantities are positively correlated, with a Pearson correlation coefficient c = 0.628 (pvalue= $1e-11$, Figure 5.8a).

This result verified the hypothesis that the heterogeneity measured by the SHscore captures the evolutionary distance of the cells belonging to the samples analyzed.

SHscore for different evolutionary spans

The SHscore was computed on the 45 pairs of samples generated from the *10Kcells* experiment, and the results were combined with those obtained in the *hom_scenario* and the *100Kcells* experiment. The samples in the three scenarios contain a comparable number of cells (~ 500) but derive from trees whose growth was stopped at a different height. This means that the sample history, in the three scenarios, diverged at different heights on the parental tree and kept on growing for a comparable number of doublings at the same mutation rate, which the generating model fixes. Therefore, sample cells, in the three different scenarios, are likely to have accumulated the same amount of heterogeneity, starting from their MRCAs, while their divergence is mainly due to the heterogeneity accumulated by their MRCAs, which are located at

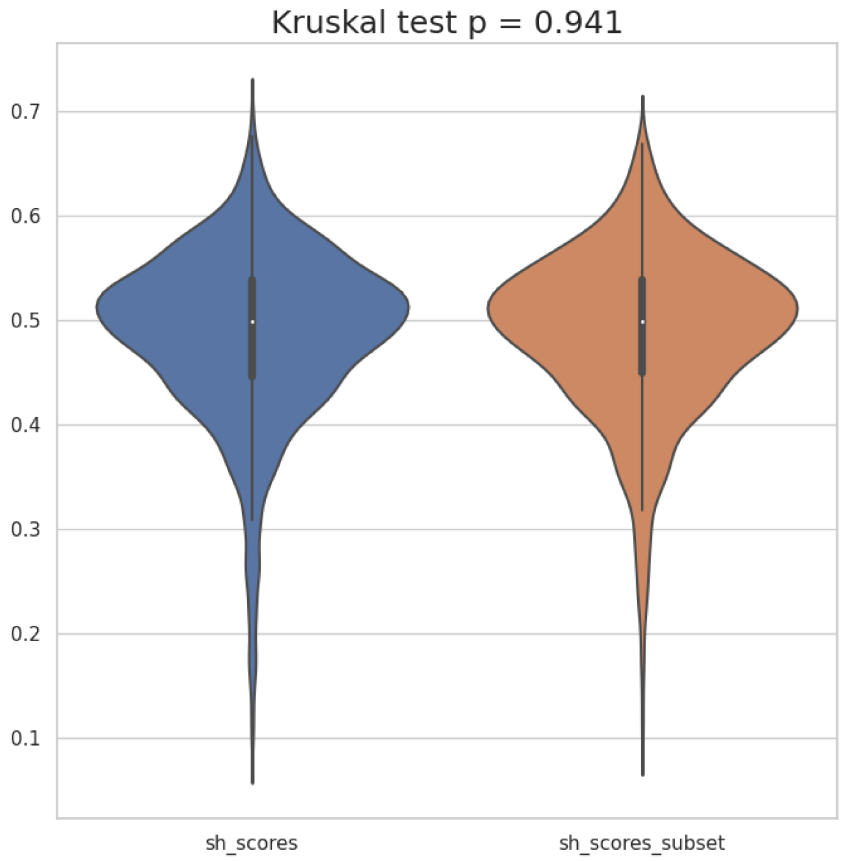


Fig. 5.7 **Supplementary Figure 2**: comparison distribution of the full set of SHscores computed for the 4950 pairs of samples and the distribution of the 1000 randomly sampled. The two set of scores are equally distributed (Kruskal-Wallis pvalue = 0.941), so the SHscore subset is representative of the full set of scores.

different distances on the parental tree (very close on 2.5K cell trees, very distant on the 100K cell tree, intermediate distance on the 10K cell tree). Figure 5.8b shows that the scores in the three scenarios are distributed around a different median (2.5K cells = median: 0.151, IQR: 0.064, 10K cells = median: 0.278, IQR: 0.061, 100K cells = median: 0.498, IQR: 0.092), which value increases as the mean distance between the MRCAs of the sample increases.

This is an additional proof of what was shown before: the closer the MRCAs, the higher the score.

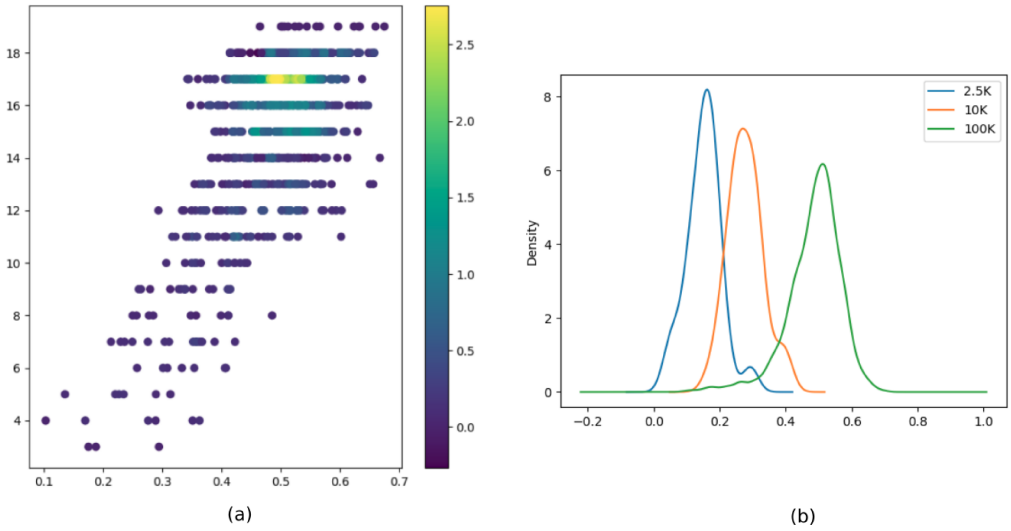


Fig. 5.8 SHscore vs Evolutionary distance. A Pearson correlation test was executed on the SHscores and the MRCA distances, demonstrating that the two quantities are positively correlated (coef = 0.628, pvalue= $1e-11$) (7a). The SHscores computed on datasets deriving from from trees which growth was stopped at a different height were grouped. The scores in the three scenarios are distributed around a different median ($2.5K$ cells = median: 0.151, IQR: 0.064, $10K$ cells = median: 0.278, IQR: 0.061, $100K$ cells = median: 0.498, IQR: 0.092), which value increases as the mean distance between the MRCAs of the sample increases (7b).

The results shown in this section allow to conclude that a score lower than 0.2 indicates that the subclones are well-mixed in the tumor sample or segregated in space, but spatial differences are so small that the tumor may be considered homogeneous. A score greater or equal to 0.2 instead suggests that different regions of the same tumors are separated by a non-negligible evolutionary distance which made them quite different and this should be considered for further analyses.

5.3.3 Experiment 3: SHscore on tumor data

This section presents three examples of application of PhyliCS on real scCNA public datasets.

Spatial subsamples from the same disease site

This example shows how PhyliCS may be used to investigate spatial intra-tumor heterogeneity at a single disease site.

PhyliCS have been used with five single-cell CNA datasets produced with Cell Ranger DNA and published on 10x Genomics website [114]. The datasets derive from five sections (`S_A`, `S_B`, `S_C`, `S_D`, `S_E`), of the same frozen breast tumor tissue and contain data related to 2137, 2224, 1722, 1916 and 2053 cells, respectively.

scCNA calling A few preliminary steps were required to produce PhyliCS input files. Specifically, 10x multi-cell alignment files were demultiplexed, using a C++ based tool, *SCtools*, developed with the SeqAn library [156]. After that, some quality checks were performed, and the CNA events were computed using Ginkgo [67]. At this point, the scCNA datasets were loaded into PhyliCS.

Data Pre-processing Using the preprocessing module, diploid or pseudo-diploid cells (ploidy ranging in the interval [1.6, 2.9]) were removed because they were considered uninformative; also, cells whose CNA profile was characterized by a high (>95th percentile) median absolute deviation (MAD) were filtered out, because they were considered noisy, due to single-cell amplification issues or ongoing DNA replication. As a result, the cells left for the five samples were 110, 1172, 1040, 1137, and 1473. Since `S_A` contained very few tumor cells compared to the other samples, it was not included in the following analysis steps.

Multi-Sample Analysis Figure 5.9a shows the graphical results produced after the aggregation phase. The cells from the four samples share a similar CNA profile and have been mixed-up by the clustering algorithm.

Figure 5.9b, instead, presents a diagram containing the SHscores computed for different sample aggregations. The value indicated as `'S_B vs. S_C vs. S_D vs. S_E'` indicates how much the samples are different from each other. According to is has been observed in the simulation experiment, the value -0.0201 indicates that the four samples show a very similar genomic make-up, which makes them almost indistinguishable. Additionally, it can be noticed that by combining the samples `S_C`, `S_D` and `S_E` and testing them against `S_B`, the SHscore grows to 0.182388, indicating that its genomic make-up may be clonally separated from that of the other samples. SHscores confirm the graphical results shown in Figure 5.9a, highlighting S_B as the more divergent sample, a result that is backed up by the

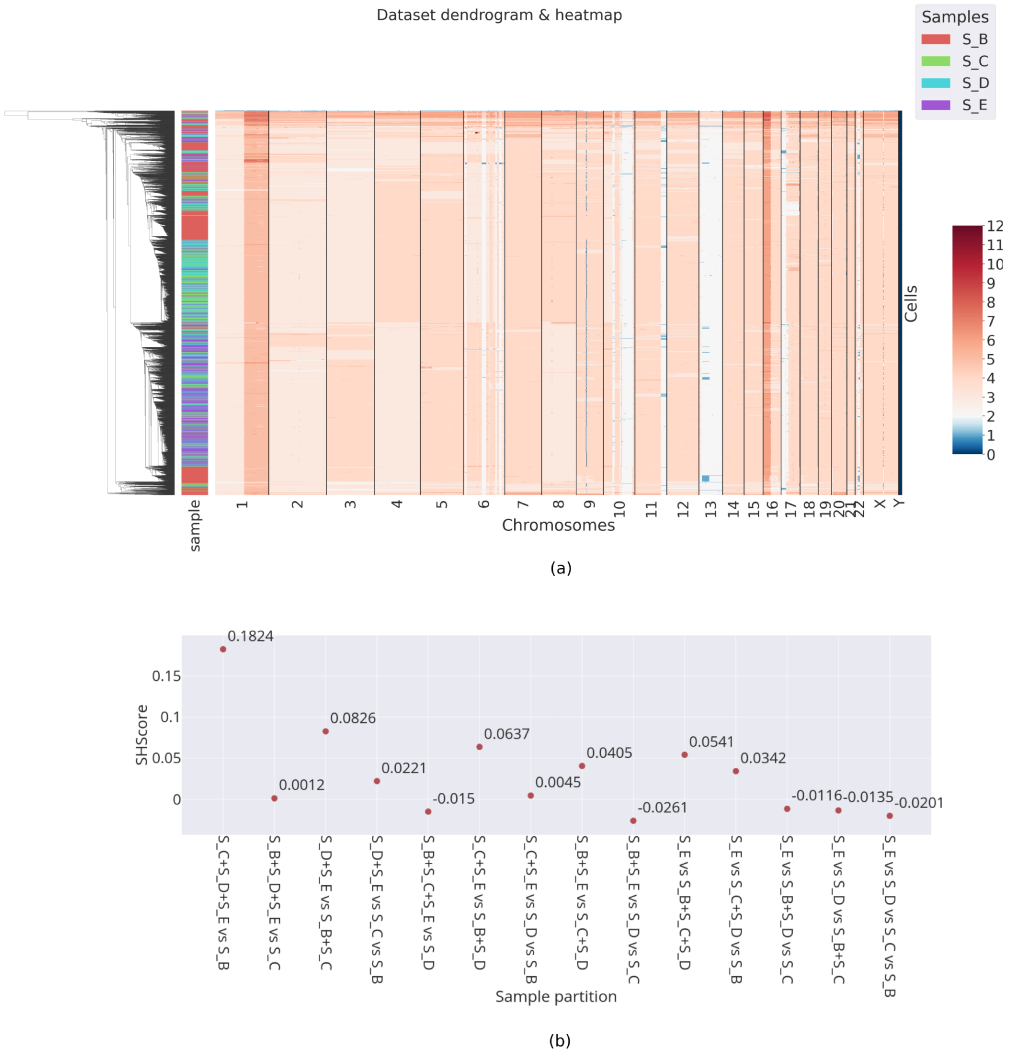


Fig. 5.9 Test case: breast tumor data. PhyliCS was teste on a scCNA dataset containing the data of five sections (S_A, S_B, S_C, S_D, S_E) of the same breast tumor. After some preliminary operations, S_A was discarded for further analysis. This experiment resulted in the evidence that the sections share similar genomic patterns (8a), with the exception of S_B; this is confirmed by the SHscore (8b), which best value (0.1824) is obtained by by aggregating S_C, S_D, S_E against S_B.

clonal reconstruction made by CHISEL [72], which reveals a subclone (J-I) that is almost private to that sample.

Spatial subsamples from the different disease sites

The proposed method has also been applied to a pair of samples derived from a primary tumor and a matched metastasis. The results of the CNA analysis performed by Garvin et al. on a dataset to validate Ginkgo [67] were used. The dataset corresponded to a primary breast tumor and its liver metastasis (T16P/M) and was used by Navin et al. [157] for their study on intra-tumor heterogeneity characterization. Since the CNA calls were available on the Ginkgo website, they could be directly loaded into PhyliCS.

Data Pre-processing Also, in this case, diploid and pseudo-diploid cells and cells with a high MAD were discarded, reducing the aggregated dataset cardinality from 100 to 42 cells.

Multi-Sample Analysis Figure 5.10 presents the results obtained from the analysis performed on this dataset. It shows that, apparently, the same cell population which initiated the tumor also seeded the metastasis, confirming the findings of the original publication [157]. The hierarchical clustering algorithm, this time, has organized cells in two separate blocks, corresponding to the two populations from the primary tumor and the metastasis. This underlines a certain degree of separation between the two samples, which the SHscore also represents. Even if it is impossible to compare scores for different sample arrangements, the SHscore (0.5361) is consistent with the results we obtained on metastatic scenarios simulations. The high SHscore means that although the primary and metastatic samples share a common mutational pattern, their following, independent evolution made them clearly distinguishable. This suggests that the differences between primary and metastatic pairs that have always been measured with bulk sequencing can be further studied with scDNA approaches [158, 159].

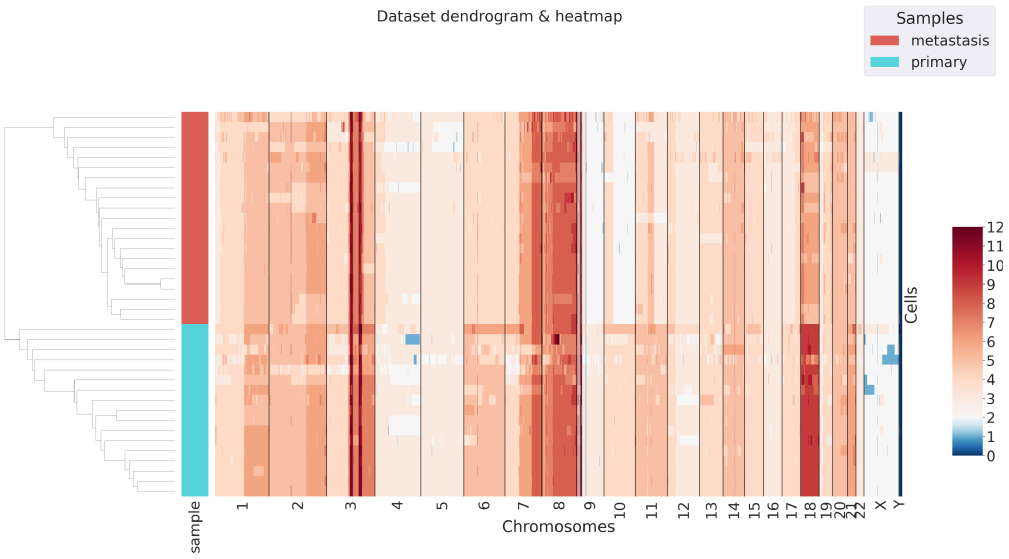


Fig. 5.10 Test case: lung tumor data. PhyliCS was tested on pair of samples derived from a primary lung tumor and a matched liver metastasis. This time, the two samples shown a certain degree of genetic diversity and where characterized by a high SHscore (0.5361).

Clonal expansion of a cell line

This example presents an extended use-case which shows how PhyliCS may be used to investigate the heterogeneity gained by a clonally expanded cell line.

In detail, a single-cell dataset, recently published by Minussi et al. [113] on NCBI Sequence Read Archive (accession number PRJNA629885) has been exploited: it contained the sequencing reads of cells from a triple-negative breast cancer cell line (MDA-MB-231) (508 cells) and those resulting from the clonal expansion of 2 single daughter cells (MDA231-EX1 and MDA231-EX2) from the parental cell line for 19 cell doublings (995 and 897 cells, respectively).

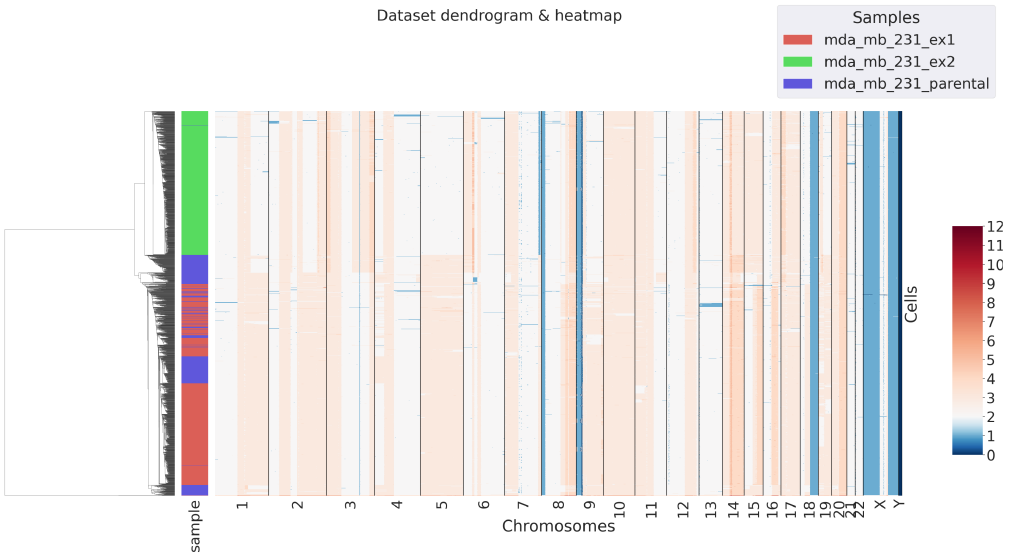
Data preparation The sequencing reads were downloaded from NCBI. Concerning the parental dataset, two batches of cells were available: 508 obtained with single-end sequencing and 312 resulting from paired-end sequencing. Only the cells sequenced using the single-end technique were kept to avoid a batch effect. The single-cell reads were aligned against the GRCh38 reference genome using BWA (v0.7.17) [160]. Low quality reads (MAPQ $<$ 20), secondary alignments, and PCR duplicates using SAMtools (v1.9) [161] were filtered out. After that, the BED files

were generated using BEDTools (v2.27.1) [162]. Finally, the CNA events were computed for the three datasets, separately, using a standalone version of Ginkgo [67], with variable binning (mean bin size = 500kb) and default options. To generate boundaries for variable-length bins for the reference genome, the method outlined by Garvin et al. [67] and implemented at the Ginkgo repository was adopted. It consists in sampling 101bp reads from the reference genome and mapping it back to itself (BWA), looking for uniquely mapping reads. After that, for the given bin size, reads are assigned to bins such that each bin has the same number of uniquely mappable reads. Consequently, intervals with higher repeat contents and low mappability will be larger than intervals with highly mappable sequences, although they will have the same number of uniquely mappable positions.

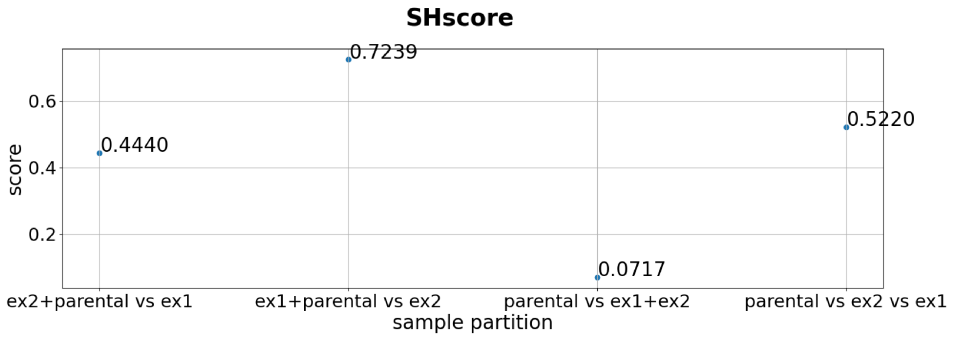
Multi-Sample Analysis The generated CNA matrices were provided to PhyliCS, and the SHscore was computed for all possible partitions of the three datasets. Figure 5.11b shows that the best SHscore (0.7102) was obtained when aggregating the MDA-MB-231-EX1 dataset with the parental one. This result indicates that MDA-MB-231-EX1 cells share a common genomic pattern with the parental cell line. This is confirmed by the results of the hierarchical clustering performed on the aggregated dataset, graphically shown in Figure 5.11a: cells from MDA-MB-231-EX1 are well mixed with the parental ones, while the cells from MDA-MB-231-EX2 are put into a completely separate block. This may be due to two reasons: the clonal expansion from MDA-MB-231-EX2 originating cell generated more heterogeneity than the other one, or the clonal subpopulation which MDA-MB-231-EX2 originating cell was sampled from is not represented in the parental dataset. Anyhow, it is possible to state that the proposed score is capable of capturing the different levels of diversity among multiple samples, and when using it comparatively, it is highly informative.

SHscore robustness to dataset cardinality MDA-MB-231 dataset was further exploited to demonstrate the SHscore robustness to dataset cardinality.

In detail, in a real-world scenario, the number of cells sequenced may vary from sample to sample. For this reason, multiple downsampling experiments were conducted on the daughter cell lines to test the robustness of the SHscore to the cardinality of the samples.



(a)



(b)

Fig. 5.11 Test case: MDA-MB-231 cell line data. PhyliCS was tested on MDA-MB-231 cell line. In details, the parental cell-line was compared to the datasets resulting from the clonal expansion of two dauther cells, MDA-MB-231-EX1 and MDA-MB-231-EX2, for 19 doublings. The datasets contained 508, 995 and 897 cells respectively. The dataset deriving from the expansion of MDA-MB-231-EX1 shown to be more similar to the parental line, with respect to the genomic profile of the data deriving from MDA-MB-231-EX2 (5.11a). In fact, the best SHscore (0.7102) was obtained when aggregating MDA-MB-231-EX1 dataset with the parental against the other one (5.11b).

The two cell lines are characterized by a comparable number of cells (995, 897) and derive from a clonal expansion of two single cells. For this reason, they are expected to be internally genetically homogeneous, and some random subsamples should not be consistently different from each other. Consequently, the SHscore is expected to remain almost stable when comparing two subsamples with varying cardinalities from the two cell lines.

As the first step, the SHscore was computed using the complete datasets. Figure 5.12 shows that the hypothesis of internal genetic homogeneity is confirmed while the SHscore value, 0.832, indicates that well distinguishable CN profiles characterize the two cell lines.

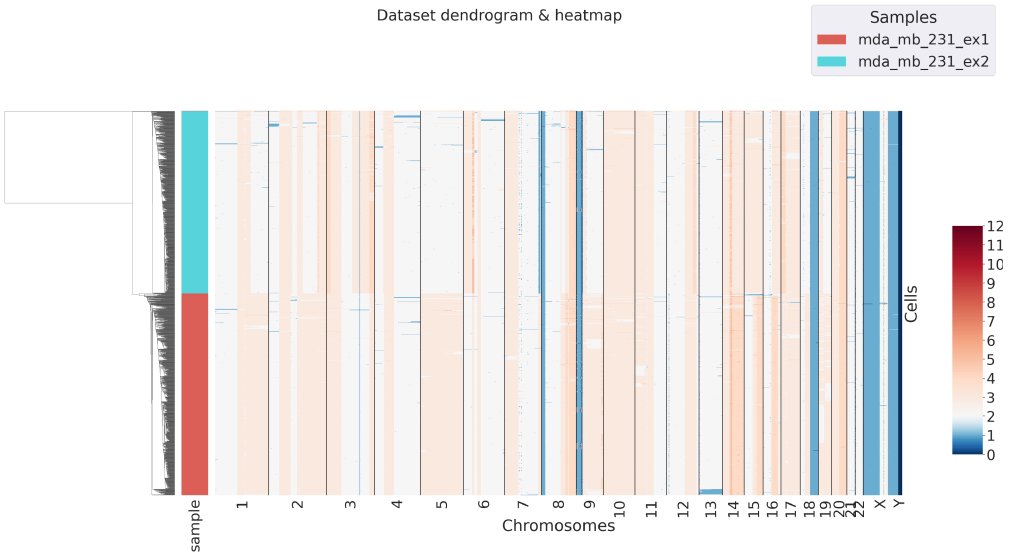


Fig. 5.12 MDA-MB-231-EX1 vs. MDA-MB-231-EX2. Multi-sample analysis was conducted on the two daughter cell lines. The hierarchical clustering algorithm separated their cells into two well-separated and internally homogeneous blocks. The SHscore (0.832106) confirmed this evidence.

After that, 9 subsamples were generated from each cell line by randomly selecting a fraction (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%) of its cells. The SHscore was computed on all pairs made of one of the two complete datasets (e.g., MDA-MB-231-EX1) and one of the subsamples from the other (e.g., MDA-MB-231-EX2_10%, MDA-MB-231-EX2_20%, etc.). As Figure 5.13 shows, the resulting SHscores fluctuated of a small quantity with respect to the initial SHscore. Specifically, they where distributed in the interval [0.815, 0.850], with median =

0.833 and IQR = 0.015. The small fluctuations of the score should not surprise because, albeit being internally homogeneous, the two cell lines still present a subclonal structure [113], so the downsampling operation may have targeted cells belonging to different subclones. Anyhow, the fact that the median result is almost identical to the original SHscore indicates that the proposed method is robust to the cardinality of the datasets and may be used to compare samples of any size; additionally, it demonstrates that the SHscore can capture heterogeneity even when the input dataset cardinalities are very unbalanced.

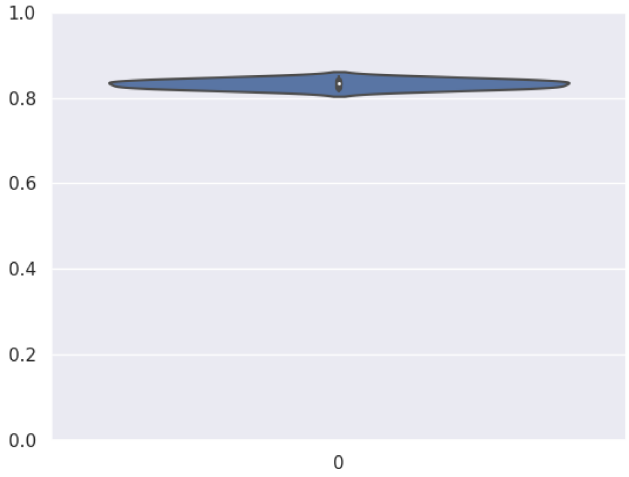


Fig. 5.13 Supplementary Figure 6: Downsampling experiment. In order to test the robustness of the proposed method both daughter cell lines were downsampled, producing 9 subsamples for both of them, each containing a fraction 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of their cells. The SHscore was computed on pairs made of one of the two full datasets against each of the subsamples of the other dataset: despite small fluctuations, due to the little amount of heterogeneity existing in the the cell lines, the SHscores (min = 0.815, max = 0.850, median = 0.833, IQR = 0.015) were comparable to that computed for the orginal dataset (0.832), confirming the proposed method is robust to the cardinality of the input datasets.

5.4 Conclusions

This chapter presented PhyliCS, a flexible and user-friendly package that allows to process scCNA calls and evaluate spatial ITH through the Spatial Heterogeneity Score. This score combines the high resolution of scDNA sequencing data and the

information provided by multi-regional sampling to indicate how much different sets of cells have diverged in their CN landscapes, allowing to get fast and easy-to-interpret information about a single tumor.

PhyliCS has been implemented as a modular and flexible Python library, with many functionalities, which guides bioinformaticians who want to explore their datasets to use a single API specific for scDNA and tailored to its analysis.

The SHscore has been tested in different scenarios. First, it was computed on 200 synthetic datasets to study its behavior in four different scenarios (spatial segregation, spatial intermixing, early metastasis spreading, and late metastasis spreading). Results obtained on this set of simulations show that SHscore correctly reflects the heterogeneity in the clonal composition of multiple samples and can therefore be used to reliably compare the heterogeneity of renal tumors with different spatial samplings available. After that, the SHscore was tested on a set of 100 simulations generated by randomly varying the mean CNA size and the mean number of gained copies, showing to be not correlated to such structural features of the CN profiles. A more extensive simulation experiment, generating two big cell-division trees, generated datasets with a significant evolutionary history. This experiment returned the evidence that the SHscore is strongly correlated to the distance between the copy-number states, which generated the cells of the samples in the analysis. This confirmed that the SHscore captures the evolutionary history of the tumor subsamples. The score was used to analyze three real scDNA datasets, reaching conclusions in agreement with state-of-the-art phylogenetic approaches [72] and the original papers [157, 113] that presented them. Finally, a downsampling experiment was conducted on two cell line data to demonstrate that the SHscore is robust to sample cardinality and may be used on unbalanced sets.

Trying to define clinically relevant thresholds for the SHscore is premature. Indeed, large cohorts of clinically annotated single-cell datasets from patients affected by different tumors would be required to correlate the evolutionary features of each tumor with its clinical characteristics and subsequently define thresholds to discriminate between “spatially segregated” and “spatially well-mixed” scenarios of clinical relevance. Unfortunately, such single-cell DNA datasets are not yet available. However, the presented extended simulation study returned the evidence that a score lower than 0.2 indicates that the subclones are well-mixed in the tumor sample or segregated in space, but spatial differences are so small that the tumor may be

considered homogeneous. A score greater or equal to 0.2 instead suggests that different regions of the same tumors are separated by a non-negligible evolutionary distance which made them quite different and this should be considered for further analyses.

One of the current limitations of PhyliCS is that all its results regarding evolutionary distances are derived from samples relationships and clustering-based metrics. This approach was adopted to draw conclusions that, albeit simplistic, are based on fewer assumptions on the mechanisms driving CN accumulation than the ones needed to perform the phylogenetic reconstruction. Being the infinite site assumption not valid for CNs, the phylogenetic reconstruction is still an open issue for single-cell data. However, it is possible to foresee that there will be more reliable methods to call SNVs on single cells in the future, opening new avenues to exploit the theoretical knowledge built upon bulk sequencing.

In summary, PhyliCS represents a valuable instrument to explore the extent of spatial heterogeneity in multi-regional tumor sampling, exploiting the potential of scCNA data.

In the future, scDNA sequencing should gain popularity, and more data will be available on public repositories; at that point, it will be possible to test and improve the score on large-scale datasets. Additionally, it will be interesting to integrate different single-cell measurements, such as ATACseq or scRNA, to extend its capabilities. The choice to develop a library should ease future endeavors in this direction.

Part II

Gene Fusion Classification

Chapter 6

Identifying The Oncogenic Potential Of Gene Fusions Exploiting miRNAs

It is estimated that oncogenic gene fusions cause about 20% of human cancer morbidity. Identifying potentially oncogenic gene fusions may improve affected patients' diagnosis and treatment.Previous approaches to this issue included exploiting specific gene-related information, such as gene function and regulation.

This chapter presents ChimerDriver, a tool to classify gene fusions as oncogenic or not oncogenic. ChimerDriver is based on a specifically designed neural network and trained on genetic and post-transcriptional information to obtain a reliable classification.

The designed neural network integrates information related to transcription factors, gene ontologies, microRNAs and other detailed information related to the functions of the genes involved in the fusion and the gene fusion structure. As a result, the performances on the test set reached 0.83 f1-score and 96% recall. The comparison with state-of-the-art tools returned comparable or higher results. Moreover, ChimerDriver performed well in a real-world case where 21 out of 24 validated gene fusion samples were detected by the gene fusion detection tool Starfusion.

ChimerDriver integrates transcriptional and post-transcriptional information in an ad-hoc designed neural network to effectively discriminate oncogenic gene fusions from passenger ones.

6.1 Scientific background

Gene fusions are one of the most common somatic mutations and are considered to be responsible for 20% of global human cancer morbidity [163, 164]. A gene fusion is a biological event where two independent genes fuse to form a hybrid gene. In the most common case, one gene retains the promoter region and the other one provides the end of the hybrid gene. The former is called 5p' gene, while the latter is called 3p' gene. The position where the break occurs is called breakpoint.

The advent of next-generation sequencing (NGS), the spread of machine and deep learning in bioinformatics [165–168] and the development of fusion detection algorithms [103, 102, 169, 170] led to the discovery of hundreds of novel fusion sequences.

However, not all gene fusions are oncogenic. Indeed, some are genuinely expressed in normal human cells [100] or constitute passenger events [101]. At the same time, other gene fusions are considered to be responsible for a significant percentage of specific kinds of tumors [171, 90, 172, 173].

A precise diagnosis of oncogenic gene fusions can inform therapeutics treatments [174, 175] and be used to predict prognosis, patient survival, and treatment response [164]. Additionally, focusing the research on a smaller number of putative oncogenic fusions a diagnosis could take less time; thus, the risks related to misdiagnosis and waiting may be significantly reduced for the patients.

However, discriminating between cancer-driver fusions and non-driver events is not a trivial task.

The first necessary step to solve this problem is performed by the fusion detection tools [103, 102, 169], that identify the candidate gene fusions relying on the sample's reads, trying to reduce as much as possible the number of false positives (i.e., detected gene fusions that are not found in the sample in later lab validation). Additional studies proposed more sophisticated approaches based on machine learning (ML) techniques applied to the output of fusion detection tools. Specifically, Oncofuse [108] and Pegasus [109] are noteworthy and use protein domains of the fusion proteins to train the models and predict the oncogenic potential of a fusion. Undoubtedly protein domains are highly informative for the characterization of gene fusions. However, using such information as a feature for the ML model requires

careful processing from scratch whenever the training database is updated with novel validated fusions.

Recently, previous works explored deep-learning (DL) techniques [176] and presented DEEPPrior [110], a DL model to perform gene fusion prioritization using amino acid sequences of the fusion proteins, based on a Convolutional Neural Network (CNN) and a bidirectional Long Short Term Memory (LSTM) network. Compared to the state-of-the-art tools, this approach is highly effective in accomplishing the classification task with the advantage of avoiding labor-intensive processing of the protein domains.

However, it is known that the oncogenic potential of a molecule depends not only on the sequence itself but also on the effect of post-transcriptional regulatory processes[177].

Transcription Factors (TFs) and micro-RNAs (miRNAs) play a decisive role in the transcriptional and post-transcriptional regulatory processes [178] and can contribute to determining the gene fusion outcome.

To date, most of the available tools exploit transcriptional information and common gene properties to accomplish this task without considering the post-transcriptional regulators affecting the oncogenic processes.

This work proposes ChimerDriver, a new DL architecture based on a Multi-Layer Perceptron (MLP) that integrates gene-related information with miRNAs and TFs, including then in the model transcriptional and post-transcriptional regulative information. Indeed, ChimerDriver exploits the knowledge about TFs and miRNAs targetting each of the genes involved in the fusion to perform gene fusion classification.

ChimerDriver was tested on multiple publicly available datasets and exhibited better classification performance with respect to the state-of-the-art tools. In the end, post-transcriptional regulators confirm the central role in discovering oncogenic processes and miRNAs; in particular, they are a precious source of information to improve the prediction of the oncogenic potential of gene fusions.

In the following, a detailed description of model, its architecture and the input datasets is provided into the Material and methods section. Results are illustrated in Results section. The discussion and conclusion are reported in Discussion and Conclusions sections, respectively.

6.2 Material and methods

This section introduces the proposed pipeline for the classification of gene fusions. In detail, after a brief overview of ChimerDriver architecture, it illustrates the classification model, the training and testing sets, and the model input features.

6.2.1 ChimerDriver architecture

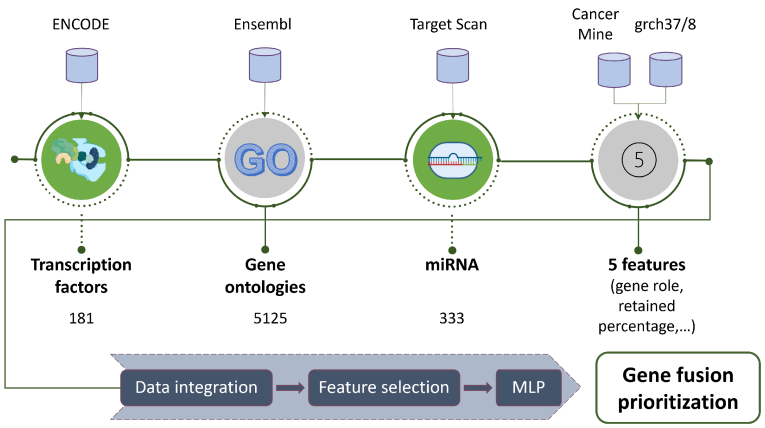


Fig. 6.1 Conceptual schema of ChimerDriver architecture.

Figure6.1 shows the conceptual schema of ChimerDriver. The tool is made of three main modules: the data integration module, which is in charge of extracting the model input features from the input data and integrating them; the feature selection model, which reduces the number of input features through a Random Forest; the classification module, which performs gene fusion prioritization using a neural network. The adopted classification model, the training and testing sets, and the input features are discussed in the following sections.

6.2.2 Model design

ChimerDriver classifies gene fusions using a Multi-Layer Perceptron (MLP). MLPs are a classical type of feed-forward neural network that, thanks to their flexibility, may be applied to multiple types of data and learn non-linear correlations among them, even when they are produced from different sources [179–181].

According to grid search results, the best network configuration was denoted by four layers with 512, 256, 128, 64 nodes, all characterized by the tanh activation function, and with a learning rate and dropout values equal to 0.01 and 0.2, respectively.

6.2.3 Dataset

The training and testing sets were carefully designed. Specifically, the training set consisted of 1765 gene fusion samples: 1059 were oncogenic, and the remaining 706 were not oncogenic. The oncogenic samples were extracted from COSMIC (Catalog of Somatic Mutations in Cancer), a popular database containing information about gene fusions involved in solid tumors and leukemias [182]. Besides, chosen oncogenic gene fusions were already experimentally validated. Finally, the 706 not-oncogenic gene fusions were reported by Babicenau et al. [183] and detected by a gene fusion detection tool in non-neoplastic tissues.

The testing set consisted of 2623 oncogenic gene fusions and 2254 not-oncogenic gene fusions. In detail, for the positive samples, the choice fell on the database provided by Gao et al. [184], which results from the application of three fusion detection tools on the TCGA database. Upon request, the authors kindly provided validated gene fusion samples, for which WGS data were available. From this collection, 2622 oncogenic gene fusions were extracted. In addition, 2254 not-oncogenic gene fusions found in healthy tissues and reported by Babicenau et al [183] were incorporated.

Finally, to avoid overfitting, the genes involved in the training set gene fusions were not present in the tested gene fusions. In this way, it was possible to verify that the model is sufficiently robust and can learn the oncogenic characteristics of the gene fusions and not specific information relating to the individual genes.

6.2.4 Input features

The model input features were selected from multiple sources to assess different gene fusions' characteristics.

The first five features are obtained from the gene fusion structure and Cancermine [185], a literature-mined database of drivers, oncogenes, and tumor suppressors

in cancer. In detail, given the breakpoint coordinates, two features correspond to the retained percentage of 5p' and 3p' genes in the gene fusion. One additional feature analyzes the strands of 5p' and 3p' genes, and it is equal to 1 if the two strands are concordant (i.e., the two genes transcribe in the same direction), 0 otherwise. The remaining two features correspond to the nature of each gene according to Cancermine [185]: 'Oncogenic', 'Driver', 'Tumor suppressor' or 'Other' when none of the above options applies. This feature contributes to assessing the functional profiling of the gene fusion.

TFs and GOs involving the fused genes were added to the aforementioned input features. In fact, multiple studies [108, 186, 187] demonstrated that using these molecules in the gene fusion classification task has a positive impact on the final model performance. Specifically, a set of 181 TFs was extracted from the ENCODE database [186], and only those related to the gene in the 5p' position were considered. Additionally, all GOs involving fused genes were selected.

Finally, all miRNAs predicted to target all 5p' and 3p' genes were included in the feature set. This information was extracted from TargetScan, a popular state-of-the-art database that predicts biological targets of miRNAs by searching for the presence of sites that match the seed region of each miRNA [188], reporting for each miRNA all possible target genes. A set of 333 miRNAs was obtained by investigating the probability of targeting the genes belonging to the gene fusion. In case of ambiguity, only the highest probability was retained. This should be the first time that post-transcriptional regulation information has been used in such a classification task.

The final feature number was 5644, which is a considerably high number compared to the number of samples in our training and test sets. Thus, we performed feature selection to reduce feature set size to avoid overfitting our dataset. The chosen feature selection method was Random Forest, by which the number of features was lowered to 310.

6.3 Results

This section discusses the results obtained with ChimerDriver and the comparison with the state-of-the-art tools. Additionally, a case study in which ChimerDriver was applied on a pair of well-known datasets is presented.

6.3.1 Results on the training set

As previously stated, ChimerDriver was trained on 1765 gene fusions, obtained from COSMIC, Catalog of Somatic Mutations in Cancer [182] and from Babicenau et al. work [183]. Given each gene fusion's breakpoint, the aforementioned features are extracted and then fed to the MLP. The model was cross-validated on the training set with the k-fold method. K value was set equal to 10. The AUC, Accuracy, F1 score, precision and recall are reported in Table 6.1. The model reached an average f1-score of 0.98 on our training set with different combinations of learning rate and dropout values.

Learn rate, dropout	AUC	Accuracy	F1	Precision	Recall
0.0001, 0.0	0.981	0.978	0.981	0.981	0.981
0.0001, 0.2	0.979	0.976	0.979	0.980	0.979
0.0001, 0.4	0.981	0.978	0.981	0.981	0.981
0.001, 0.0	0.980	0.976	0.980	0.980	0.980
0.001, 0.2	0.976	0.972	0.975	0.968	0.984
0.001, 0.4	0.977	0.974	0.977	0.980	0.975
0.01, 0.0	0.982	0.979	0.982	0.986	0.979
0.01, 0.2	0.982	0.979	0.982	0.983	0.982
0.01, 0.4	0.980	0.976	0.980	0.983	0.978

Table 6.1 Cross validation results with the k-fold method. The value of k was set equal to 10.

6.3.2 Results on the test set

The model was tested on 4877 gene fusions. 2623 oncogenic gene fusions were retrieved from the work of Gao et al.[184] and the remaining 2254 were gene fusions found in healthy tissues and reported by Babicenau et al. [183]. The test samples

were entirely independent from the training samples. The model returned a 0.83 f1-score and 96% recall when tested on this set of gene fusions.

6.3.3 miRNA impact on the classification performance

The miRNA features were extracted from TargetScan [188], a popular database that maps gene-miRNA pairs providing various kinds of information, included the miRNA's probability of targeting the specific gene during post-transcriptional regulation. This value was extracted for both 5p' and 3p' genes and it is intended to represent the involvement of miRNAs in gene fusion processes. Figure 6.2 highlights the impact of the miRNA features in the classification by displaying the confusion matrices including and excluding miRNAs from the evaluation. The impact of miRNAs is particularly evident when looking at the number of false-negative gene fusions, which is almost doubled when miRNAs are not considered. Including miRNAs in the classification task increases the recall value from 93% to 96%.

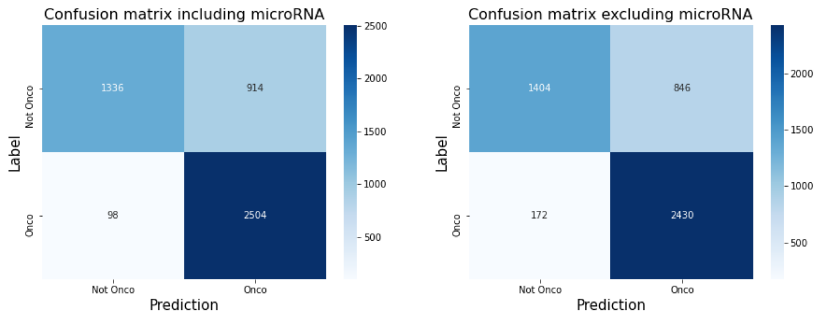


Fig. 6.2 Confusion matrices reporting the MLP results including miRNAs (on the left) and excluding miRNA features (on the right).

6.3.4 Comparison with state of the art

ChimerDriver performances were compared to those reported by three related works: Oncofuse [108], DEEPrior [110], and Pegasus [109]. To compare the results in the most unbiased way, the experimental conditions of the three tools were reproduced and ChimerDriver was applied.

Oncofuse

To test the robustness of the proposed method, the training set and testing set used by Oncofuse [108] were retrieved and used to train and test the proposed model.

Oncofuse training samples were extracted from TICDB [189], a curated database that contains gene fusions found in tumor samples, and from a collection of fusion genes [190], and read-through transcripts [191] found in normal cells named NORM-RTH. Oncofuse's authors then built the oncogenic testing set by merging oncogenic gene fusions from CHIMERDB [192] and NGS, respectively oncogenic fusions predicted by gene fusion detection tools and fusions discovered and published in NGS studies about cancer [193–196]. On the other hand, not-oncogenic testing samples were taken from Refseq [197] and CGC [198], two databases that report unbroken gene fusions. In particular, the samples that belong to CGC involve unbroken oncogenic genes.

All the previously listed features (see Material and methods for details) were processed and gathered, except for the two features related to the retained percentage of genes since the provided dataset omitted the breakpoint information.

The ChimerDriver model was tailored to this comparison. In detail, obtained 281 input features: the strands and the involvement in oncogenic processes of both 5p' and 3p' genes, 93 TFs, 155 miRNAs, and 30 GOs. The maximum number of epochs was set to 50, and the number of nodes per layer was 256, 128, 64, and 32 (the associated activation functions were the relu, sigmoid, relu, and sigmoid, respectively). The learning rate was fixed to 0.03, while the dropout value applied to each layer was 0.4.

Figure 6.3 shows the comparison of the classification results obtained by ChimerDriver and Oncofuse. Precisely, the green bars correspond to the results achieved by Oncofuse, as reported by its authors [108], while the blue ones show the results obtained by ChimerDriver. Similar to Oncofuse paper, the results are displayed separately for each database. The bar diagram shows the percentage of driver gene fusions detected by the model. As it can be noticed, when trained and tested on the samples provided by Oncofuse, ChimerDriver provided better results with respect to those illustrated in the original paper.

Specifically, as reported by Figure 6.3, 95% of TICDB samples were correctly classified as driver gene fusions by ChimerDriver as opposed to the assumed 90%

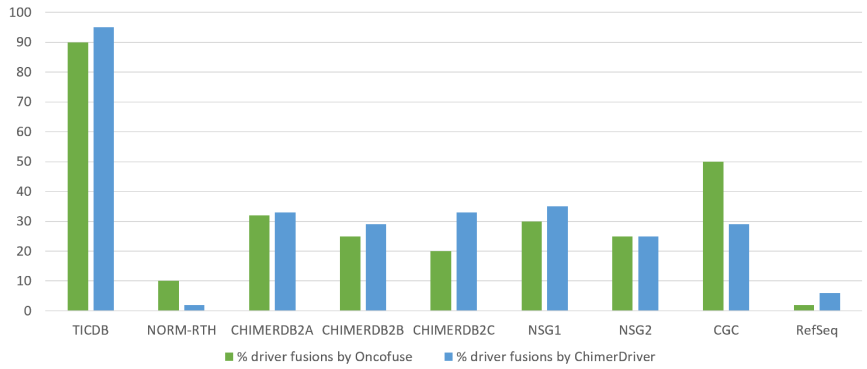


Fig. 6.3 The green bars correspond to the results reported by Shugay M. et al. in their paper. In blue the results obtained by ChimerDriver are displayed.

reported by Oncofuse, furthermore 2% of the NORM-RTH samples were incorrectly classified as driver gene fusions by ChimerDriver as opposed to the assumed 10% reported by Oncofuse.

ChimerDriver successfully outperformed Oncofuse in the oncogenic gene fusion databases used as a test set, namely ChimerDB2A, ChimerDB2B, ChimerDB2C, NGS1 and NGS2. ChimerDriver identified more or a comparable amount of oncogenic gene fusions in each database with respect to Oncofuse, correctly classifying about 1/3 of the samples.

ChimerDriver minimized the number of detected driver fusions of unbroken oncogenic genes, identifying a lower number of driver gene fusions in CGC database, as additional test set.

When tested on the not-oncogenic samples in RefSeq database, Chimerdriver returned a slightly higher number of driver gene fusions.

In general, it is possible to conclude that even without the information on the retained percentage of genes, ChimerDriver outperformed Oncofuse in the great majority of cases.

DEEPrior

DEEPrior is a DL-based classifier that performs gene prioritization using protein sequences obtained from the gene fusion samples. Its architecture is based on a CNN and an LSTM network. It was trained on a dataset extracted from COSMIC [182],

and Babicenau et al.'s study [183] and tested on the part of the oncogenic gene fusion collection validated by Gao et al. [184]. DEEPrior reconstructs the protein sequences from gene fusion breakpoint information and assigns to each gene fusions an oncogenic score defining its oncogenic probability. Gene fusions are ordered according to the oncogenic score and highly scored fusions are prioritized as drivers. In this sense, DEEPrior main aim consists in providing a reliable classification prediction (oncogenic or not) according to the oncogenic score.

ChimerDriver was trained and tested on the DEEPrior training set and test set (*Dataset 2* in DEEPrior paper). As a result, ChimerDriver correctly classified 96% of oncogenic gene fusions from the test set. On the contrary, DEEPrior prioritized as driver the 32.48% of gene fusions found in the test set. Since DEEPrior aims at classifying only highly probable oncogenic fusions, the percentage of prioritized gene fusions is not directly comparable with the classification performances obtained with ChimerDriver. ChimerDriver provides a classification result for each gene fusion, while DEEPrior classifies a tiny percentage of gene fusions in the dataset.

It is possible conclude that the ChimerDriver approach exploits different sources of information (TFs, GOs, miRNAs) while DEEPrior focuses on identifying the oncogenic potential of a gene fusion through its protein sequence without considering the effect of post-transcriptional regulators.

At the same time, ChimerDriver ensures a less computationally intensive approach in the training phase than DEEPrior.

Pegasus

To further assess ChimerDriver classification performances, its performance was compared to those of Pegasus [109], a state-of-the-art tool for gene fusion detection and classification purposes. Pegasus exploits a traditional machine learning model to predict of driver gene fusion, namely a gradient tree boosting algorithm.

Also, in this case, ChimerDriver was trained and tested on the gene fusion samples used to develop and validate Pegasus.

The training dataset was strongly unbalanced towards the negative samples, comprising over 9923 negative samples out of 10162 gene fusions. Not to penalize

the MLP architecture, which is particularly sensitive to class unbalance, the number of negative gene fusions was lowered to 239, namely the number of positive samples.

ChimerDriver was cross-validated on 10 folds using the aforementioned training samples. It should be noted that, as a result of balancing the classes, the model was given a fairly small number of training examples. In the end, the f1-score was equal to 0.89 with a learning rate and dropout, respectively equal to 0.001 and 0.

Pegasus' test set accounted for 78 gene fusions, 39 oncogenic and 39 not oncogenic, respectively. According to Pegasus authors, the curated subset of 39 oncogenic gene fusions was almost entirely correctly classified by Pegasus, which reported 0.97 of AUC and 0.95 of AUC for the not oncogenic samples.

Pegasus intently selected as negative examples 39 not oncogenic gene fusions containing at least a tumor suppressor or an oncogene. The rationale is that these gene fusions would be most challenging for a classification task. ChimerDriver correctly classified 27 out of the 39 not-oncogenic gene fusions enforcing the notion that the model can generalize even on not oncogenic gene fusions. On the other hand, the oncogenic test samples represented a more difficult classification task for ChimerDriver, which detected 17 oncogenic gene fusions. It should be noted that ChimerDriver model was initially trained and tested on a wide variety of gene fusions proving its ability to learn and generalize well when given a fair amount of examples. On the contrary, since Pegasus was developed and refined on particular tissues, a reduced number of samples is used as a training set.

It is highly probable that the small number of samples in the Pegasus training set negatively impacted the ChimerDriver training phase, which benefits from a wider number of gene fusions. Therefore, ChimerDriver performances, when trained and tested on Pegasus datasets, are negatively affected, hindering the likelihood of reaching the outcome reported by the Pegasus authors.

Table 6.2 summarizes the results presented in this section.

6.3.5 Case study

Finally, to assess ChimerDriver's performances in a clinical context, two well-known studies were selected: 6 breast cancer samples [199] and 4 prostate cancer

[1]Percentage of highly probable oncogenic gene fusions (prioritization)

Test set	ChimerDriver	OncoFuse
CHIMERDB2A	33%	32%
CHIMERDB2B	29%	25%
CHIMERDB2C	33%	20%
NGS1	35%	30%
NGS2	25%	25%
CGC	29%	50%
RefSeq	6%	2%
		DEEPrior
DEEPrior test set	96%	32.48%[1]
		Pegasus
Pegasus test set	56%	97.43%

Table 6.2 ChimerDriver vs state-of-the-art tools. ChimerDriver performances compared to those reported by three related works: Oncofuse, DEEPrior, and Pegasus.

samples [200] in which 24 gene fusions are reported to be experimentally validated. The samples are all RNA-seq data. They were processed with STAR-fusion [104] to identify which gene fusions were found in these samples by a standard and accurate fusion detection tool. 21 out of the 24 validated gene fusions were detected with STAR-fusion and subsequently processed with ChimerDriver to confirm the ability to detect oncogenic gene fusions in a real-world case correctly. Figure 6.4 shows the results of this assessment. Specifically, the gene fusions marked in gray were not detected by STAR-fusion hence were not available to ChimerDriver for further processing. The training dataset and the training parameters are described in detail in the Material and methods section like the ones generally used in the ChimerDriver training procedure. On the 21 samples, ChimerDriver wrongly classified as not oncogenic the three oncogenic gene fusions marked in orange. By inspecting the oncogenic role of 5p' and 3p' genes and the retained percentage in the gene fusion, a possible explanation for the wrong classification could be hypothesized. Concerning the ACACA-STAC2 gene fusion, no information on the involvement of any of the two genes was provided to the algorithm. So, although most of the portion of both genes was retained after the gene fusion event, ChimerDriver was probably unsure about their role in oncogenic processes. As for the GLB1-CMTM7 fusion, the algorithm was aware that the latter gene is involved in tumor suppression; on the other hand, the retained percentage of CMTM7 was less than 45%. This probably led the network to conclude that there was not enough gene left in the gene fusion

to cause issues. Similarly, in the CPNE1-PI3 fusion, the percentage of retained genes (respectively 25% and 40%) was probably too low to label the gene fusion as oncogenic even if the genes were associated with the roles oncogenic and driver, respectively. Finally, ChimerDriver correctly classified the 18 remaining gene fusions as oncogenic. Hence, ChimerDriver correctly classified 18 out of 21 oncogenic gene fusions, demonstrating that the specifically designed neural network is proficient in learning and generalizing from a consistent number of gene fusion samples. Moreover, the information gathered from the different sources and provided to the tool as features proved to be particularly effective in discerning oncogenic and not-oncogenic fusions even in a realistic circumstance.

Validated gene fusions	Detected by ChimerDriver
ANKHD1_PCDH1	Yes
CCDC85C_SETD3	Yes
WDR67_ZNF704	Not detected by Starfusion
CYTH1_EIF3H	Yes
DHX35_ITCH	Yes
BSG_NFIX	Yes
PPP1R12A_SEPT10	Yes
NOTCH1_NUP214	Yes
BCAS4_BCAS3	Yes
ARFGEF2_SULF2	Yes
RPS6KB1_TMEM49	Not detected by Starfusion
TMPRSS2_ERG	Yes

Validated gene fusions	Detected by ChimerDriver
ACACA_STAC2	No
RPS6KB1_SNF8	Yes
VAPB_IKZF3	Yes
ZMYND8_CEP25	Not detected by Starfusion
RAB22A_MYO9B	Yes
SKA2_MYO19	Yes
STARD3_DOK5	Yes
LAMP1_MCF2L	Yes
GLB1_CMTM7	No
CPNE1_PI3	No
TATDN1_GSDMB	Yes
RARA_PKIA	Yes

Fig. 6.4 The 24 oncogenic gene fusions validated in prostate and breast tumor samples are reported. STAR-fusion did not detect the three gene fusions marked in gray hence were not available to ChimerDriver for further processing. ChimerDriver correctly classified as oncogenic 18 out of the 21 available gene fusions.

6.4 Discussion

Identifying oncogenic gene fusions is of crucial importance in cancer detection and prognosis. To date, state-of-the-art tools exploit transcriptional and GOs information without considering the post-transcriptional regulators in predicting the oncogenic potential of a gene fusion. Here, ChimerDriver was introduced, a novel tool to accomplish the aforementioned task exploiting transcriptional and post-transcriptional regulators. In detail, ChimerDriver focuses on miRNAs post-transcriptional effect as a key feature to perform the prediction.

ChimerDriver is based on an ad-hoc designed neural network embedding miRNAs, transcription factors, gene ontologies, and gene-specific information to predict gene fusions' oncogenic potential. The model is stable and exhibits excellent classification performance (f1-score = 0.98).

The classifier was tested against three state-of-the-art tools: Oncofuse, DEEPrior, and Pegasus.

With respect to Oncofuse, post-transcriptional regulation was introduced to perform the classification and, as a result, ChimerDriver outperformed Oncofuse in the great majority of tested cases.

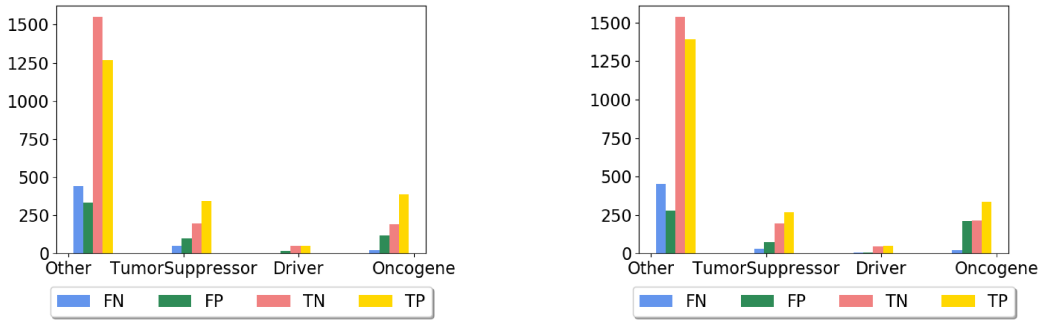
In particular, ChimerDriver performed better than Oncofuse on the test set, correctly classifying as oncogenic about 1/3 of the oncogenic gene fusions. ChimerDriver identified a comparable or higher amount of oncogenic gene fusions outperforming Oncofuse results in each positive test case. ChimerDriver minimized the number of detected driver fusions in 'unbroken oncogenic genes' (negative testing samples) extracted from CGC compared to Oncofuse. This result confirmed the ability of ChimerDriver in generalizing and taking advantage of the given set of features to make a correct prediction. As previously presented in the Results section about Pegasus comparison, this statement is true even when the samples contain an oncogene or a tumor suppressor. ChimerDriver returned a slightly higher number of oncogenic gene fusions than Oncofuse when tested on the RefSeq database of 'unbroken not-oncogenic genes'. The breakpoint information was not available in Oncofuse datasets. Therefore, to perform an unbiased comparison with Oncofuse, the breakpoint information was neglected by the ChimerDriver model. Consequently, the percentage of driver gene fusions detected by ChimerDriver on RefSeq was slightly higher than expected, probably because the tool could not profit from the breakpoint information.

ChimerDriver also outperformed DEEPrior in terms of the number of classified gene fusion. In particular, ChimerDriver correctly identified 96% of oncogenic gene fusions in the dataset used to test DEEPrior, which prioritized as oncogenic only 32.48% of the samples. It should be noted that the goals of DEEPrior and ChimerDriver are slightly different. The first prioritizes gene fusions, returning those with an oncogenic probability greater than a threshold (typically 80%). ChimerDriver performs an immediate classification of each gene fusion by integrat-

ing transcriptional and post-transcriptional features in the assessment. The outcome of ChimerDriver is remarkable in terms of the number of oncogenic samples that were correctly classified while also enlightening because it stresses the extent to which miRNAs are involved in the oncogenic processes of gene fusions.

Moreover, the performances of ChimerDriver were compared to the ones reported by Pegasus authors. According to their research, the latter could correctly classify almost all of the test samples. After training and testing ChimerDriver on the gene fusions provided by the authors, it was observed that the number of detected oncogenic samples was lower than the results reported by Pegasus. As already stated in the Results section, the number of training samples was lowered in order to balance the oncogenic and not oncogenic classes. However, the limited number of samples processed by ChimerDriver in the training phase has probably inhibited the neural network from learning efficiently. In addition, Pegasus's authors specify that the negative validation samples included at least one oncogene or tumor suppressor. Remarkably, to make a prediction, ChimerDriver also relies on the role of each gene in oncogenic processes (e.g., driver, oncogene, or tumor suppressor), making the classification task particularly arduous to tackle. In addition, Pegasus and, consequently, ChimerDriver were trained on a reduced number of samples, thus impacting ChimerDriver performances. Nevertheless, ChimerDriver correctly classified most of the not oncogenic gene fusions enforcing the notion that the model can generalize well in this situation.

This work focused on the integration of information coming from different databases to improve the current state-of-the-art research on classifying oncogenic gene fusions. Additionally, a neural network was designed explicitly for this task. However, the main contribution of the present work is the introduction of miRNAs in the classification model. In fact, despite miRNAs role in determining the oncogenic potential of gene fusions has been demonstrated, they had never been considered in such a task. The present study showed that they could significantly improve the model performance. In particular, they halved the number of false negatives and improved the recall of the model. It is possible to conclude that miRNAs, being involved in the regulation of gene fusion-related protein, are a promising indicator of the oncogenic potential of gene fusions.



(a) (b)

Fig. 6.5 Distribution of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) regarding Cancermine information for both 5p' and 3p' genes(respectively Figure 6.5a) and 6.5b)). Noticeably, FPs are never tumor suppressors, drivers or oncogenes.

The main limitation of the proposed method is that some gene fusions are misclassified. To better investigate ChimerDriver classification with respect to the Cancermine [185] role, the distribution of the Cancermine roles (e.g. tumor suppressor, driver, oncogene, other) for 5p' gene (Figure 6.5a) and 3p' gene (Figure 6.5b) was analized. In addition, test set samples are divided in each role according to the classification results (false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN)). TP samples are characterized for 5p' ad 3p' genes by a prevalence of suppressors and oncogenes. On the contrary, TN mostly refer to the 'other' CancerMine role. Consequently, FP samples could consist of oncogenes (in particular for 3p' gene) and FN samples are hardly ever related to tumor suppressors, drivers, or oncogenes. In this sense, FP and FN samples reflect ChimerDriver behavior on TP and TN, respectively. In a clinical context, FN misclassified samples are unlikely to be tested for in-lab validation since most involve genes with no specific oncogene/tumor suppressor role. FP samples instead would have been considered for experimental validation, which would exclude them from oncogenic fusions. However, laboratories would still benefit from a selection of putative oncogenic gene fusions.

6.5 Conclusions

Gene fusions are a common mutation that is nowadays known to be responsible for about 1/5 of human cancers. It is of the uttermost importance to correctly

identify gene fusions to improve cancer detection and prognosis. Considering that the state-of-the-art tools exploit transcriptional and gene information neglecting post-transcriptional regulations, this work established the value of miRNAs in achieving superior classification performances.

To conclude, this chapter presented ChimerDriver, a novel and stable DL architecture based on a Multi-Layer Perceptron (MLP), that, for the first time, combines gene-level features with TFs and miRNAs targetting the gene fusion to perform its classification and prioritization.

ChimerDriver was trained and tested on a consistent number of gene fusions. The final results highlight the impact of miRNAs in evaluating the oncogenic potential of gene fusions. It is possible to infer that the inclusion of miRNAs represents a valuable advantage in identifying oncogenic gene fusions.

ChimerDriver can become a valuable tool for research laboratories to predict the oncogenic potential of gene fusions. Indeed, the expensive validations could be targeted cost-effectively with this easy-to-use tool; additionally, it may speed up identifying novel and potentially oncogenic gene fusions, allowing for better diagnosis, classification, and treatment of cancer patients.

Chapter 7

Conclusions

The advent of next-generation sequencing (NGS) technologies led to a rapid increase in the amount and complexity of genomic sequencing data. New computational techniques are required to exploit such an abundance of data and extract helpful information. This thesis focused on the study and the application of statistical and AI-based techniques on human cancer sequencing data.

The first part of this thesis is dedicated to statistical and machine learning methods to model intra-tumor heterogeneity. In particular, it explored how single-cell CNA data may be used to characterize and quantify intra-tumor genetic diversity.

The first contribution has been the design of a pipeline capable of producing multi-sample copy-number aberrations analysis on large-scale single-cell DNA sequencing data and performing a qualitative inspection of spatial and temporal tumor heterogeneity. The proposed pipeline, albeit simple, overcomes the limits of closed source software, giving on the one hand researchers the possibility to explore the complete breadth of their data and, on the other, a fully-fledged and easy to run pipeline to obtain multi-sample and heterogeneity visualizations.

The second result consists of a formal and systematic assessment of well-known clustering methods to study their performance and identify the approach providing the most accurate reconstruction of phylogenetic relationships. This study demonstrated that the algorithms that do not require to be seeded with the cluster number outperformed the others. Specifically, Affinity Propagation outperformed the others when no dimensionality reduction was performed, while density-based algorithms had outstanding results on top of PCA and UMAP results.

Finally, the third and uttermost contribution, in this context, has been the design and development of PhyliCS, the first tool which exploits scCNA data from multiple samples from the same tumor to estimate whether the different clones of a tumor are well mixed or spatially separated, through a specifically designed score, the SHscore. This score combines the high resolution of scDNA sequencing data and the information provided by multi-regional sampling to indicate how much different sets of cells have diverged in their CN landscapes, allowing to get fast and easy-to-interpret information about a single tumor. The SHscore has been evaluated in a variety of simulation settings. Results show that the proposed score accurately represents heterogeneity in the clonal structure of multiple samples and indirectly reflects the evolutionary history of tumor subsamples. Additionally, having been developed as a modular and flexible Python library, PhyliCS provides many functionalities which guide bioinformaticians who want to explore their datasets to use a single API specific for scDNA and tailored to its analysis.

Considering the effectiveness of AI tools in resolving complicated biological issues characterized by a scarcity of domain expertise, they were also used to explore the carcinogenicity of gene fusions. Specifically, the second part of this thesis consists of developing deep learning-based methods to classify gene fusions as oncogenic or not-oncogenic. In this regard, the primary contributions have been the development of a neural network which, for the first time, combines gene-level features with TFs and miRNAs targetting the gene fusion to perform its classification and prioritization. The neural network was trained and tested on a consistent number of gene fusions. The final results highlight that the incorporation of miRNAs in the classificatio model halved the number of false negatives and improved the recall of the model. Considering that the state-of-the-art tools exploit transcriptional and gene information neglecting post-transcriptional regulations, this work established the value of miRNAs in achieving superior classification performances.

7.1 Global considerations

This thesis presented new methods and applications to exploit the abundance of NGS data to deconvolve cancer complexity.

In particular, they represent a valuable instrument for life scientists interested in modeling the processes underlying cancer evolution and leading to inter and

intra-tumor heterogeneity. They also aim to improve the discovery of biological hallmarks to classify tumors in subtypes, stratify patients, predict response to drugs, and improve precision medicine. In the future, they may also be used in the clinical context.

The proposed methods adopt new approaches to address some well-known biological questions. In particular, it has been shown how single-cell DNA sequencing, combined with multi-regional sampling, may help understand the intrinsic complexity of tumors and how multiple heterogeneous genetic features may be integrated to understand the nature of gene fusions.

In the future, more data will be available on public repositories; therefore, it will be possible to investigate further the biological phenomena addressed by this thesis and to continue to improve and validate the presented computational methods.

To conclude, this thesis provided life scientists with some powerful instruments to better investigate genomic cancer data and produce some advancements in improving the diagnosis, classification, and treatment of cancer patients.

References

[1] El Mustapha Bahassi and Peter J Stambrook. Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, 29(5):303–310, 2014.

[2] Tom Ronan, Zhijie Qi, and Kristen M Naegle. Avoiding common pitfalls when clustering biological data. *Science signaling*, 9(432):re6–re6, 2016.

[3] Emilie Baro, Samuel Degoul, Régis Beuscart, and Emmanuel Chazard. Toward a literature-driven definition of big data in healthcare. *BioMed research international*, 2015, 2015.

[4] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz Jr, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[5] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: a cancer journal for clinicians*, 2022.

[6] Stefania Morganti, Paolo Tarantino, Emanuela Ferraro, Paolo D'Amico, Bruno Achutti Duso, and Giuseppe Curigliano. Next generation sequencing (ngs): a revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics*, pages 9–30, 2019.

[7] Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[8] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2022. Last accessed February 9, 2022.

[9] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[10] National Cancer Institute. Definition of whole-genome sequencing – NCI Dictionary of Cancer Terms, 2022. Last accessed February 9, 2022.

[11] Illumina. Precise analysis of DNA–protein binding sequences, 2022. Last accessed February 9, 2022.

[12] Illumina. Study gene expression using RNA sequencing, 2022. Last accessed February 9, 2022.

[13] Illumina. A rapid, sensitive method for profiling accessible chromatin across the genome, 2022. Last accessed February 9, 2022.

[14] James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature methods*, 11(1):25–27, 2014.

[15] Fábio CP Navarro, Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson, and Mark Gerstein. Genomics and data science: an application within an umbrella. *Genome biology*, 20(1):1–11, 2019.

[16] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[17] NHGRI. Genomic Data Science, 2022. Last accessed February 9, 2022.

[18] IBM Cloud Education. Artificial Intelligence (AI), 2022. Last accessed February 9, 2022.

[19] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M Krawitz, Susanne B Kamphausen, Martin Zenker, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, 25(1):60–64, 2019.

[20] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8):1161–1170, 2018.

[21] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.

[22] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A Graham, Guido Sanguinetti, and Andrea Sottoriva. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods*, 15(9):707–714, 2018.

[23] Yana Bromberg, Guy Yachdav, and Burkhard Rost. Snap predicts effect of mutations on protein function. *Bioinformatics*, 24(20):2397–2398, 2008.

[24] Carles Ferrer-Costa, Josep Lluis Gelpí, Leire Zamakola, Ivan Parraga, Xavier De La Cruz, and Modesto Orozco. Pmut: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, 21(14):3176–3178, 2005.

[25] Hongjian Qi, Chen Chen, Haicang Zhang, John J Long, Wendy K Chung, Yongtao Guan, and Yufeng Shen. Mvp: predicting pathogenicity of missense variants by deep learning. *BioRxiv*, page 259390, 2018.

[26] Daniel Quang, Yifei Chen, and Xiaohui Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 2015.

[27] Richard J Dobson, Patricia B Munroe, Mark J Caulfield, and Mansoor AS Saqi. Predicting deleterious nssnps: an analysis of sequence and structural attributes. *BMC bioinformatics*, 7(1):1–9, 2006.

[28] Vidhya Gomathi Krishnan and David R Westhead. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, 19(17):2199–2209, 2003.

[29] Lei Bao and Yan Cui. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, 21(10):2185–2190, 2005.

[30] Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667, 2009.

[31] Emidio Capriotti, Leonardo Arbiza, Rita Casadio, Joaquín Dopazo, Hernán Dopazo, and Marc A Marti-Renom. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Human Mutation*, 29(1):198–204, 2008.

[32] Jia Xu, Pengwei Yang, Shang Xue, Bhuvan Sharma, Marta Sanchez-Martin, Fang Wang, Kirk A Beaty, Elinor Dehan, and Baiju Parikh. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Human genetics*, 138(2):109–124, 2019.

[33] Georg Bartsch Jr, Anirban P Mitra, Sheetal A Mitra, Arpit A Almal, Kenneth E Steven, Donald G Skinner, David W Fry, Peter F Lenehan, William P Worzel, and Richard J Cote. Use of artificial intelligence and machine learning algorithms with gene expression profiling to predict recurrent nonmuscle invasive urothelial carcinoma of the bladder. *The Journal of urology*, 195(2):493–498, 2016.

[34] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[35] Paul Sajda. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.*, 8:537–565, 2006.

[36] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[37] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[38] Marco Gerlinger and Charles Swanton. How darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *British journal of cancer*, 103(8):1139–1143, 2010.

[39] Timothy A Yap, Marco Gerlinger, P Andrew Futreal, Lajos Pusztai, and Charles Swanton. Intratumor heterogeneity: seeing the wood for the trees. *Science translational medicine*, 4(127):127ps10–127ps10, 2012.

[40] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.

[41] Rebecca A Burrell and Charles Swanton. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular oncology*, 8(6):1095–1111, 2014.

[42] Carlo C Maley, Patricia C Galipeau, Jennifer C Finley, V Jon Wongsurawat, Xiaohong Li, Carissa A Sanchez, Thomas G Paulson, Patricia L Blount, Rosa-Ana Risques, Peter S Rabinovitch, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, 38(4):468–473, 2006.

[43] Thomas J Lynch, Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A Okimoto, Brian W Brannigan, Patricia L Harris, Sara M Haserlat, Jeffrey G Supko, Frank G Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

[44] Simone Zaccaria and Benjamin J Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature communications*, 11(1):1–13, 2020.

[45] Matteo Manica, Hyunjae Ryan Kim, Roland Mathis, Philippe Chouvarine, Dorothea Rutishauser, Laura De Vargas Roditi, Bence Szalai, Ulrich Wagner, Kathrin Oehl, Karim Saba, et al. Inferring clonal composition from multiple tumor biopsies. *NPJ systems biology and applications*, 6(1):1–13, 2020.

[46] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[47] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, et al. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, 2014.

[48] Marleen M Nieboer, Lambert CJ Dorssers, Roy Straver, Leendert HJ Looijenga, and Jeroen de Ridder. Targetclone: A multi-sample approach for reconstructing subclonal evolution of tumors. *PloS one*, 13(11):e0208002, 2018.

[49] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

[50] Yao Xiao, Xueqing Wang, Hongjiu Zhang, Peter J Ulintz, Hongyang Li, and Yuanfang Guan. Fastclone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. *Nature communications*, 11(1):1–11, 2020.

[51] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[52] Stefan C Dentro, David C Wedge, and Peter Van Loo. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor perspectives in medicine*, 7(8):a026625, 2017.

[53] Jan Schröder, Arthur Hsu, Samantha E Boyle, Geoff Macintyre, Marek Cmero, Richard W Tothill, Ricky W Johnstone, Mark Shackleton, and Anthony T Papenfuss. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8):1064–1072, 2014.

[54] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17):e165–e165, 2013.

[55] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):1–16, 2014.

[56] Roland F Schwarz, Anne Trinh, Botond Sipos, James D Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput Biol*, 10(4):e1003535, 2014.

[57] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):1–20, 2015.

[58] Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1–16, 2015.

[59] Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.

[60] Eugene Urrutia, Hao Chen, Zilu Zhou, Nancy R Zhang, and Yuchao Jiang. Integrative pipeline for profiling dna copy number and inferring tumor phylogeny. *Bioinformatics*, 34(12):2126–2128, 2018.

[61] Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):1–12, 2019.

[62] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.

[63] Steven A McCarroll and David M Altshuler. Copy-number variation and association studies of human disease. *Nature genetics*, 39(7):S37–S42, 2007.

[64] Noemi Andor, Billy T Lau, Claudia Catalanotti, Vijay Kumar, Anuja Sathe, Kamila Belhocine, Tobias D Wheeler, Andrew D Price, Maengseok Song, Zeljko Dzakula, et al. Joint single cell dna-seq and rna-seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *BioRxiv*, page 445932, 2020.

[65] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, 14(2):167, 2017.

[66] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, et al. Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221, 2019.

[67] Tyler Garvin, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz. Interactive analysis and assessment of single-cell copy-number variations. *Nature methods*, 12(11):1058, 2015.

[68] Bjorn Bakker, Aaron Taudt, Mirjam E Belderbos, David Porubsky, Diana CJ Spierings, Tristan V de Jong, Nancy Halsema, Hinke G Kazemier, Karina Hoekstra-Wakker, Allan Bradley, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome biology*, 17(1):1–15, 2016.

[69] Xuefeng Wang, Hao Chen, and Nancy R Zhang. Dna copy number profiling using single-cell sequencing. *Briefings in bioinformatics*, 19(5):731–736, 2018.

[70] Xiao Dong, Lei Zhang, Xiaoxiao Hao, Tao Wang, and Jan Vijg. Sccnv: a software tool for identifying copy number variation from single-cell whole-genome sequencing. *Frontiers in genetics*, 11, 2020.

[71] Rujin Wang, Dan-Yu Lin, and Yuchao Jiang. Scope: a normalization and copy-number estimation method for single-cell dna sequencing. *Cell Systems*, 10(5):445–452, 2020.

[72] Simone Zaccaria and Benjamin J Raphael. Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature biotechnology*, pages 1–8, 2020.

[73] Mengyuan Li, Zhilan Zhang, Lin Li, and Xiaosheng Wang. An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. *Communications biology*, 3(1):1–19, 2020.

[74] Nadine Norton, Pooja P Advani, Daniel J Serie, Xochiquetzal J Geiger, Brian M Necela, Bianca C Axenfeld, Jennifer M Kachergus, Ryan W Feathers, Jennifer M Carr, Julia E Crook, et al. Assessment of tumor heterogeneity, as evidenced by gene expression profiles, pathway activation, and gene copy number, in patients with multifocal invasive lobular breast tumors. *PloS one*, 11(4):e0153411, 2016.

[75] Won-Chul Lee, Lixia Diao, Jing Wang, Jianhua Zhang, Emily B Roarty, Susan Varghese, Chi-Wan Chow, Junya Fujimoto, Carmen Behrens, Tina Cascone, et al. Multiregion gene expression profiling reveals heterogeneity in molecular subtypes and immunotherapy response signatures in lung cancer. *Modern Pathology*, 31(6):947–955, 2018.

[76] Youngjune Park, Sangsoo Lim, Jin-Wu Nam, and Sun Kim. Measuring intratumor heterogeneity by network entropy using rna-seq data. *Scientific reports*, 6(1):1–12, 2016.

[77] An Hoai Truong, Viktoriia Sharmanska, Clara Limback-Stanic, and Matthew Grech-Sollars. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neuro-Oncology Advances*, 2(1):vdaa110, 2020.

[78] National Cancer Institute. Definition of gene fusion – NCI Dictionary of Cancer Terms, 2022. Last accessed February 14, 2022.

[79] National Cancer Institute. Definition of Philadelphia chromosome – NCI Dictionary of Cancer Terms, 2022. Last accessed April 8, 2022.

[80] Janet D Rowley. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293, 1973.

[81] George Manolov and Yanka Manolova. Marker band in one chromosome 14 from burkitt lymphomas. *Nature*, 237(5349):33–34, 1972.

[82] Emma Shtivelman, Batia Lifshitz, Robert P Gale, and Eli Canaani. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature*, 315(6020):550–554, 1985.

[83] Orna Dreazen, Ivana Klisak, Gary Jones, Winston G Ho, Robert S Sparkes, and Robert Peter Gale. Multiple molecular abnormalities in ph1 chromosome positive acute lymphoblastic leukaemia. *British journal of haematology*, 67(3):319–324, 1987.

[84] B Lifshitz, E Fainstein, C Marcelle, E Shtivelman, R Amson, RP Gale, and Eli Canaani. bcr genes and transcripts. *Oncogene*, 2(2):113–117, 1988.

[85] Brittany C Parker, Matti J Annala, David E Cogdell, Kirsi J Granberg, Yan Sun, Ping Ji, Xia Li, Joy Gumin, Hong Zheng, Limei Hu, et al. The tumorigenic fgfr3-tacc3 gene fusion escapes mir-99a regulation in glioblastoma. *The Journal of clinical investigation*, 123(2), 2013.

[86] Devendra Singh, Joseph Minhow Chan, Pietro Zoppoli, Francesco Niola, Ryan Sullivan, Angelica Castano, Eric Minwei Liu, Jonathan Reichel, Paola Porrati, Serena Pellegatta, et al. Transforming fusions of fgfr and tacc genes in human glioblastoma. *Science*, 337(6099):1231–1235, 2012.

[87] Sarah V Williams, Carolyn D Hurst, and Margaret A Knowles. Oncogenic fgfr3 gene fusions in bladder cancer. *Human molecular genetics*, 22(4):795–803, 2013.

[88] Yi-Mi Wu, Fengyun Su, Shanker Kalyana-Sundaram, Nickolay Khazanov, Bushra Ateeq, Xuhong Cao, Robert J Lonigro, Pankaj Vats, Rui Wang, Su-Fang Lin, et al. Identification of targetable fgfr gene fusions in diverse cancers. *Cancer discovery*, 3(6):636–647, 2013.

[89] Eugen C Minca, Bryce P Portier, Zhen Wang, Christopher Lanigan, Carol F Farver, Yan Feng, Patrick C Ma, Valeria A Arrossi, Nathan A Pennell, and Raymond R Tubbs. Alk status testing in non–small cell lung carcinoma: Correlation between ultrasensitive ihc and fish. *The Journal of molecular diagnostics*, 15(3):341–346, 2013.

[90] Fredrik Mertens, Cristina R Antonescu, and Felix Mitelman. Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes. *Genes, Chromosomes and Cancer*, 55(4):291–310, 2016.

[91] Yasuhito Arai, Yasushi Totoki, Fumie Hosoda, Tomoki Shirota, Natsuko Hama, Hiromi Nakamura, Hidenori Ojima, Koh Furuta, Kazuaki Shimada, Takuji Okusaka, et al. Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma. *Hepatology*, 59(4):1427–1434, 2014.

[92] Rui Wang, Haichuan Hu, Yunjian Pan, Yuan Li, Ting Ye, Chenguang Li, Xiaoyang Luo, Lei Wang, Hang Li, Yang Zhang, et al. Ret fusions define a unique molecular and clinicopathologic subtype of non–small-cell lung cancer. *Journal of clinical oncology*, 30(35):4352–4359, 2012.

[93] Brian V Balgobind, Susana C Raimondi, Jochen Harbott, Martin Zimmermann, Todd A Alonzo, Anne Auvrignon, H Berna Beverloo, Myron Chang, Ursula Creutzig, Michael N Dworzak, et al. Novel prognostic subgroups in childhood 11q23/mll-rearranged acute myeloid leukemia: results of an international retrospective study. *Blood, The Journal of the American Society of Hematology*, 114(12):2489–2496, 2009.

[94] Stephen X Skapek, James Anderson, Frederic G Barr, Julia A Bridge, Julie M Gastier-Foster, David M Parham, Erin R Rudzinski, Timothy Triche, and Douglas S Hawkins. Pax-foxo1 fusion status drives unfavorable outcome for children with rhabdomyosarcoma: a children's oncology group report. *Pediatric blood & cancer*, 60(9):1411–1417, 2013.

[95] John A Liu Yin, Michelle A O'Brien, Robert K Hills, Sarah B Daly, Keith Wheatley, and Alan K Burnett. Minimal residual disease monitoring by quantitative rt-pcr in core binding factor aml allows risk stratification and predicts relapse: results of the united kingdom mrc aml-15 trial. *Blood, The Journal of the American Society of Hematology*, 120(14):2826–2835, 2012.

[96] Nicolas Duployez, Olivier Nibourel, Alice Marceau-Renaut, Christophe Willekens, Nathalie Helevaut, Aurélie Caillault, Céline Villenet, Karine Celli-Lebras, Nicolas Boissel, Eric Jourdan, et al. Minimal residual disease monitoring in t (8; 21) acute myeloid leukemia based on runx1-runx1t1 fusion quantification on genomic dna. *American journal of Hematology*, 89(6):610–615, 2014.

[97] Ryo Tamura, Kosuke Yoshihara, Kaoru Yamawaki, Kazuaki Suda, Tatsuya Ishiguro, Sosuke Adachi, Shujiro Okuda, Ituro Inoue, Roel GW Verhaak, and

Takayuki Enomoto. Novel kinase fusion transcripts found in endometrial cancer. *Scientific reports*, 5(1):1–9, 2015.

[98] Brittany C Parker, Manon Engels, Matti Annala, and Wei Zhang. Emergence of fgfr family gene fusions as therapeutic targets in a wide spectrum of solid tumours. *The Journal of pathology*, 232(1):4–15, 2014.

[99] Felix Y Feng, J Chad Brenner, Maha Hussain, and Arul M Chinnaiyan. Molecular pathways: targeting ets gene fusions in cancer. *Clinical Cancer Research*, 20(17):4442–4448, 2014.

[100] Milana Frenkel-Morgenstern, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Del Pozo, Michael Tress, Rory Johnson, Roderic Guigo, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric rna transcripts. *Genome research*, 22(7):1231–1242, 2012.

[101] Nicolas Stransky, Ethan Cerami, Stefanie Schalm, Joseph L Kim, and Christoph Lengauer. The landscape of kinase fusions in cancer. *Nature communications*, 5(1):1–10, 2014.

[102] Matthew K Iyer, Arul M Chinnaiyan, and Christopher A Maher. Chimerascan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011.

[103] Andrew McPherson, Fereydoun Hormozdiari, Abdalnasser Zayed, Ryan Giuliany, Gavin Ha, Mark GF Sun, Malachi Griffith, Alireza Heravi Moussavi, Janine Senz, Nataliya Melnyk, et al. defuse: an algorithm for gene fusion discovery in tumor rna-seq data. *PLoS computational biology*, 7(5):e1001138, 2011.

[104] Brian J Haas, Alexander Dobin, Bo Li, Nicolas Stransky, Nathalie Pochet, and Aviv Regev. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome biology*, 20(1):1–16, 2019.

[105] Daniel Nicorici, Mihaela Şatalan, Henrik Edgren, Sara Kangaspeska, Astrid Murumägi, Olli Kallioniemi, Sami Virtanen, and Olavi Kilkku. Fusioncatcher–a tool for finding somatic fusion genes in paired-end rna-sequencing data. *biorxiv*, page 011650, 2014.

[106] Daehwan Kim and Steven L Salzberg. Tophat-fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology*, 12(8):1–15, 2011.

[107] Jikun Wu, Wenqian Zhang, Songbo Huang, Zengquan He, Yanbing Cheng, Jun Wang, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. Soapfusion: a robust and effective computational fusion discovery tool for rna-seq reads. *Bioinformatics*, 29(23):2971–2978, 2013.

[108] Mikhail, Shugay and Iñigo, Ortiz de Mendíbil and José L, Vizmanos and Francisco J, Novo. "oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions". *Bioinformatics*, 29(20):2539–46, 10 2013.

[109] Abate Francesco et al. Pegasus: A Comprehensive Annotation and Prediction Tool for Detection of Driver Gene Fusions in Cancer. *BMC systems biology*, 2014.

[110] Marta Lovino, Maria Serena Ciaburri, Gianvito Urgese, Santa Di Cataldo, and Elisa Ficarra. DEEPrior: a deep learning tool for the prioritization of gene fusions. *Bioinformatics*, 36(10):3248–3250, 02 2020.

[111] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

[112] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90, 2011.

[113] Darlan C Minussi, Michael D Nicholson, Hanghui Ye, Alexander Davis, Kaile Wang, Toby Baker, Maxime Tarabichi, Emi Sei, Haowei Du, Mashiat Rabbani, et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, 592(7853):302–308, 2021.

[114] 10x Genomics. 10x Genomics: Biology at True Resolution, 2019. Last accessed May 8, 2019.

[115] Xiao Dong, Lei Zhang, Xiaoxiao Hao, Tao Wang, and Jan Vijg. SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing. *bioRxiv*, 2019.

[116] Michael A Ortega, Olivier Poirion, Xun Zhu, Sijia Huang, Thomas K Wolfgruber, Robert Sebra, and Lana X Garmire. Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clinical and translational medicine*, 6(1):46, 2017.

[117] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[118] Knut Reinert, Temesgen Hailemariam Dadi, Marcel Ehrhardt, Hannes Hauswedell, Svenja Mehringer, René Rahn, Jongkyu Kim, Christopher Pockrandt, Jörg Winkler, Enrico Siragusa, Gianvito Urgese, and David Weese. The seqan c++ template library for efficient sequence analysis: A resource for programmers. *Journal of biotechnology*, 261:157–168, 2017.

[119] Xiaohui Ni, Minglei Zhuo, Zhe Su, Jianchun Duan, Yan Gao, Zhijie Wang, Chenghang Zong, Hua Bai, Alec R Chapman, Jun Zhao, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proceedings of the National Academy of Sciences*, 110(52):21083–21088, 2013.

[120] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1):105–117, 2010.

[121] 10x Genomics. What is Cell Ranger DNA?, 2020. Last accessed June 26, 2020.

[122] Mario Köppen. The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, volume 1, pages 4–8, 2000.

[123] Xin-Sheng Hu, Yang Hu, and Xiaoyang Chen. Testing neutrality at copy-number-variable loci under the finite-allele and finite-site models. *Theoretical Population Biology*, 112:1–13, 2016.

[124] Giovanni Iacono, Elisabetta Mereu, Amy Guillaumet-Adkins, Roser Corominas, Ivon Cuscó, Gustavo Rodríguez-Esteban, Marta Gut, Luis Alberto Pérez-Jurado, Ivo Gut, and Holger Heyn. bigscale: an analytical framework for big-scale single-cell data. *Genome research*, 28(6):878–890, 2018.

[125] Xian Fan, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Benchmarking tools for copy number aberration detection from single-cell dna sequencing data. *bioRxiv*, page 696179, 2019.

[126] Michael GB Blum and Olivier François. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006.

[127] Ruli Gao, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature genetics*, 48(10):1119, 2016.

[128] Joseph Felsenstein, J Archie, W Day, W Maddison, C Meacham, F Rohlf, and D Swofford. The newick tree format, 1986.

[129] Metin Balaban, Niema Moshiri, Uyen Mai, Xingfan Jia, and Siavash Mirarab. Treecluster: Clustering biological sequences using phylogenetic trees. *PloS one*, 14(8):e0221068, 2019.

[130] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[131] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[132] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

[133] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[134] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE, 2017.

[135] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[136] Renato Cordeiro De Amorim and Christian Hennig. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324:126–145, 2015.

[137] Pedro R Peres-Neto, Donald A Jackson, and Keith M Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4):974–997, 2005.

[138] 10x Genomics. Isolation of Nuclei for Single Cell DNA Sequencing, 2020. Last accessed June 26, 2020.

[139] 10x Genomics. USER GUIDE. Chromium Single Cell DNA Reagent Kits, 2020. Last accessed June 26, 2020.

[140] Chunhua Du, Jie Yang, Qiang Wu, and Feng Li. Integrating affinity propagation clustering method with linear discriminant analysis for face recognition. *Optical Engineering*, 46(11):110501, 2007.

[141] Darong Lai and Hongtao Lu. Identification of community structure in complex networks using affinity propagation clustering method. *Modern Physics Letters B*, 22(16):1547–1566, 2008.

[142] Guojun Gan and Michael Kwok-Po Ng. Subspace clustering using affinity propagation. *Pattern Recognition*, 48(4):1455–1464, 2015.

[143] Jianjun Liu and Jianquan Kan. Recognition of genetically modified product based on affinity propagation clustering and terahertz spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 194:14–20, 2018.

[144] Osama Abu Abbas. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, 5(3), 2008.

[145] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[146] Leland McInnes. Using UMAP for Clustering, 2020. Last accessed June 26, 2020.

[147] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[148] Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81, 2018.

[149] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[150] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[151] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[152] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[153] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[154] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

[155] Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Assessing the performance of methods for copy number aberration detection from single-cell dna sequencing data. *PLoS computational biology*, 16(7):e1008012, 2020.

[156] Knut Reinert, Temesgen Hailemariam Dadi, Marcel Ehrhardt, Hannes Hauswedell, Svenja Mehringer, René Rahn, Jongkyu Kim, Christopher Pockrandt, Jörg Winkler, Enrico Siragusa, Gianvito Urgese, and David Weesee. The seqan c++ template library for efficient sequence analysis: A resource for programmers. *Journal of biotechnology*, 261:157–168, 2017.

[157] Nicholas Navin, Alexander Krasnitz, Linda Rodgers, Kerry Cook, Jennifer Meth, Jude Kendall, Michael Riggs, Yvonne Eberling, Jennifer Troge, Vladimir Grubor, et al. Inferring tumor progression from genomic heterogeneity. *Genome research*, 20(1):68–80, 2010.

[158] Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E Navin. Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research*, 27(8):1287–1299, 2017.

[159] Devon A Lawson, Kai Kessenbrock, Ryan T Davis, Nicholas Pervolarakis, and Zena Werb. Tumour heterogeneity and metastasis at single-cell resolution. *Nature cell biology*, 20(12):1349–1360, 2018.

[160] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[161] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *GigaScience*, 10(2):giab008, 2021.

[162] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[163] Mridula Nambiar, Vijayalakshmi Kari, and Sathees C Raghavan. Chromosomal translocations in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1786(2):139–152, 2008.

[164] Mitelman Felix, Johansson Bertil, and Mertens Fredrik. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7:233–245, 2007.

[165] Marta Lovino, Gianpaolo Bontempo, Giansalvo Cirrincione, and Elisa Ficarra. Multi-omics classification on kidney samples exploiting uncertainty-aware models. In *International Conference on Intelligent Computing*, pages 32–42. Springer, 2020.

[166] Marta Lovino, Vincenzo Randazzo, Gabriele Ciravegna, Pietro Barbiero, Elisa Ficarra, and Giansalvo Cirrincione. A survey on data integration for multi-omics sample clustering. *Neurocomputing*, 2021.

[167] Ilaria Roberti, Marta Lovino, Santa Di Cataldo, Elisa Ficarra, and Gianvito Urgese. Exploiting gene expression profiles for the automated prediction of connectivity between brain regions. *International journal of molecular sciences*, 20(8):2035, 2019.

[168] Pietro Barbiero, Marta Lovino, Mattia Siviero, Gabriele Ciravegna, Vincenzo Randazzo, Elisa Ficarra, and Giansalvo Cirrincione. Unsupervised multi-omic data fusion: The neural graph learning network. In *International Conference on Intelligent Computing*, pages 172–182. Springer, 2020.

[169] Brian J. Haas, Alex Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas G. Doak, Nathalie Pochet, Jing Sun, Catherine J. Wu, Thomas R. Gingeras, and Aviv Regev. Star-fusion: Fast and accurate fusion transcript detection from rna-seq. *bioRxiv*, 2017.

[170] Heyer Erin E. et al. Diagnosis of fusion genes using targeted RNA sequencing. *Nature Communications*, 2019.

[171] M Natividad Lobato, Markus Metzler, Lesley Drynan, Alan Forster, Richard Pannell, and Terence H Rabbitts. Modeling chromosomal translocations using conditional alleles to recapitulate initiating events in human leukemias. *Journal of the National Cancer Institute Monographs*, 2008(39):58–63, 2008.

[172] Scott A. Tomlins, Bharathi Laxman, Sooryanarayana Varambally, Xuhong Cao, Jindan Yu, Beth E. Helgeson, Qi Cao, John R. Prensner, Mark A. Rubin, Rajal B. Shah, Rohit Mehra, and Arul M. Chinnaiyan. Role of the tmprss2-erg gene fusion in prostate cancer. *Neoplasia*, 10(2):177 – IN9, 2008.

[173] Kalpana Kannan, Cristian Coarfa, Pei-Wen Chao, Liming Luo, Yan Wang, Amy E. Brinegar, Shannon M. Hawkins, Aleksandar Milosavljevic, Martin M. Matzuk, and Laising Yen. Recurrent bcam-akt2 fusion gene leads to a constitutively activated akt2 fusion kinase in high-grade serous ovarian carcinoma. *Proceedings of the National Academy of Sciences*, 112(11):E1272–E1277, 2015.

[174] Brian J Druker. Imatinib as a paradigm of targeted therapies. *Advances in cancer research*, 91(1):1–30, 2004.

[175] Alice T Shaw, Beow Y Yeap, Benjamin J Solomon, Gregory J Riely, Justin Gainor, Jeffrey A Engelman, Geoffrey I Shapiro, Daniel B Costa, Sai-Hong I Ou, Mohit Butaney, et al. Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring alk gene rearrangement: a retrospective analysis. *The lancet oncology*, 12(11):1004–1012, 2011.

[176] Marta Lovino, Gianvito Urgese, Enrico Macii, Santa Di Cataldo, and Elisa Ficarra. A deep learning approach to the screening of oncogenic gene fusions in humans. *International journal of molecular sciences*, 20(7):1645, 2019.

[177] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by micrornas: are the answers in sight? *Nature reviews genetics*, 9(2):102–114, 2008.

[178] Natalia J Martinez and Albertha JM Walhout. The interplay between transcription factors and micrornas in genome-scale regulatory networks. *Bioessays*, 31(4):435–445, 2009.

[179] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[180] S.K. Pal and S. Mitra. Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, 3(5):683–697, 1992.

[181] Dennis W Ruck, Steven K Rogers, Matthew Kabrisky, Mark E Oxley, and Bruce W Suter. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE transactions on neural networks*, 1(4):296–298, 1990.

[182] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, 39(suppl_1):D945–D950, 2010.

[183] Mihaela Babiceanu, Fujun Qin, Zhongqiu Xie, Yuemeng Jia, Kevin Lopez, Nick Janus, Loryn Facemire, Shailesh Kumar, Yuwei Pang, Yanjun Qi, et al. Recurrent chimeric fusion rnas in non-cancer tissues and cells. *Nucleic acids research*, 44(6):2859–2872, 2016.

[184] Qingsong Gao, Wen-Wei Liang, Steven M Foltz, Gnanavel Mutharasu, Reyka G Jayasinghe, Song Cao, Wen-Wei Liao, Sheila M Reynolds, Matthew A Wyczalkowski, Lijun Yao, Lihua Yu, Ken Sun, Sam Q ; Fusion Analysis Working Group ; Cancer Genome Atlas Research Network ; Chen, Alexander J Lazar, Ryan C Fields, Michael C Wendl, Brian A Van Tine, Ravi Vij, Feng Chen, Matti Nykter, Shmulevich Ilya, and Li Ding. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep*, 2018.

[185] Lever Jake, Zhao Eric Y., Grewal Jasleen, Jones Martin R., and Jones Steven J. M. Cancermine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nature Methods*, 16:505–507, 2019.

[186] ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–40, 10 2004.

[187] Andrew D Yates, Premanand Achuthan, Akanni, et al. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, 11 2019.

[188] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P. Bartel. Predicting effective microrna target sites in mammalian mrnas. *eLife*, 4:e05005, aug 2015.

[189] Francisco J Novo, Inigo Ortiz de Menibil, and José L Vizmanos. Ticdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC genomics*, 8(1):1–5, 2007.

[190] Milana Frenkel-Morgenstern, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Del Pozo, Michael Tress, Rory Johnson, Roderic Guigo, et al. Chimeras taking shape: potential functions of

proteins encoded by chimeric rna transcripts. *Genome research*, 22(7):1231–1242, 2012.

[191] Serban Nacu, Wenlin Yuan, Zhengyan Kan, Deepali Bhatt, Celina Sanchez Rivers, Jeremy Stinson, Brock A Peters, Zora Modrusan, Kenneth Jung, Somasekar Seshagiri, et al. Deep rna sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC medical genomics*, 4(1):1–22, 2011.

[192] Pora Kim, Suhyeon Yoon, Namshin Kim, Sanghyun Lee, Minjeong Ko, Haeseung Lee, Hyunjung Kang, Jaesang Kim, and Sanghyuk Lee. Chimerdb 2.0—a knowledgebase for fusion genes updated. *Nucleic acids research*, 38(suppl_1):D81–D85, 2010.

[193] Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, and Kristine Kleivi. Inga rye, sandra nyberg, maija wolf, anne lise borresen dale et olli kallioniemi: Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biology*, 12(1):R6, 2011.

[194] Onur Sakarya, Heinz Breu, Milan Radovich, Yongzhi Chen, Yulei N Wang, Catalin Barbacioru, Sowmi Utiramerur, Penn P Whitley, Joel P Brockman, Paolo Vatta, et al. Rna-seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comput Biol*, 8(4):e1002464, 2012.

[195] Yan W Asmann, Brian M Necela, Krishna R Kalari, Asif Hossain, Tiffany R Baker, Jennifer M Carr, Caroline Davis, Julie E Getz, Galen Hostetter, Xing Li, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer research*, 72(8):1921–1928, 2012.

[196] Matteo Benelli, Chiara Pescucci, Giuseppina Marseglia, Marco Severgnini, Francesca Torricelli, and Alberto Magi. Discovering chimeric transcripts in paired-end rna-seq data by using ericscript. *Bioinformatics*, 28(24):3232–3239, 2012.

[197] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[198] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[199] Henrik Edgren, Astrid Murumagi, Sara Kangaspeska, Daniel Nicorici, Vesa Hongisto, Kristine Kleivi, Inga H Rye, Sandra Nyberg, Maija Wolf, Anne-Lise Borresen-Dale, et al. Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome biology*, 12(1):1–13, 2011.

[200] Chunxiao Wu, Alexander W Wyatt, Andrew McPherson, Dong Lin, Brian J McConeghy, Fan Mo, Robert Shukin, Anna V Lapuk, Steven J M. Jones, Yongjun Zhao, et al. Poly-gene fusion transcripts and chromothripsis in prostate cancer. *Genes, Chromosomes and Cancer*, 51(12):1144–1153, 2012.

Appendix A

List of the published works

This Ph.D. thesis presents the main research activities and personal contributions to the scientific community in recent years.

These works have been presented through the following publications:

- Journal papers:
  - Lovino, M., Montemurro, M., Barrese, V. S., & Ficarra, E. (2022). Identifying the oncogenic potential of gene fusions exploiting miRNAs. Journal of Biomedical Informatics, 104057.
  - Montemurro M, Grassi E., Pizzino C. G, Bertotti A., Ficarra E. & Urgese G. (2021). PhyliCS: a Python library to explore scCNA data and quantify spatial tumor heterogeneity. Journal paper in *BMC Bioinformatics*, vol. 22, pp. 1-21. ISSN 1471-2105
- Conference proceedings:
  - Montemurro M, Urgese G., Grassi E., Pizzino C. G, Bertotti A. & Ficarra E. (2020). Effective evaluation of clustering algorithms on single-cell CNA data. Conference paper in *ICBBE 2020 - 7th International Conference on Biomedical and Bioinformatics Engineering*, Kyoto, 06-09 November 2020. ISBN: 978-1-4503-8822-1
  - Montemurro M, Grassi E., Urgese G., Pizzino C. G, Bertotti A. & Ficarra E. (2019). Single-cell DNA Sequencing Data: a Pipeline for Multi-

Sample Analysis. Abstract in *ACM Celebration of Women in Computing: womENcourage 2019*, Roma, 16-18 September 2019.

– Montemurro M, Grassi E., Urgese G., Parisi E., Pizzino C. G, Bertotti A. & Ficarra E. (2019). Single-cell DNA Sequencing Data: a Pipeline for Multi-Sample Analysis. Conference paper in *CIBB 2019 - 16th International Conference on Computational Intelligence in Bioinformatics and Biostatistics*, Bergamo, 4-6 September 2019