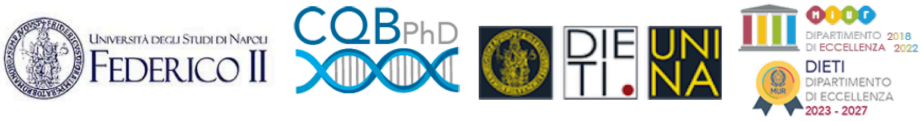


Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVI Cycle

Thesis for the Degree of Doctor of Philosophy

AI-based computational methods for tumor heterogeneity characterization, early cancer detection and oncology drug target

by
Antonio De Falco

Advisor: Prof. Michele Ceccarelli

Co-advisor: Prof. Luigi Cerulo

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

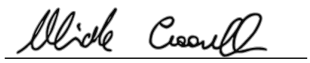
AI-based computational methods for tumor heterogeneity characterization, early cancer detection and oncology drug target

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Computational And Quantitative Biology
by

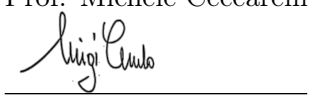
Antonio De Falco

October 2023

Approved as to style and content by



Prof. Michele Ceccarelli , Advisor



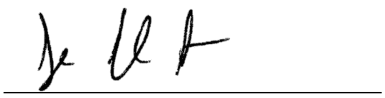
Prof. Luigi Cerulo, Co-advisor

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Computational And Quantitative

Biology is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, February 23, 2024



Antonio De Falco

Abstract

The Ph.D. thesis presents a comprehensive analysis framework that exploits the advances in sequencing technologies and statistical machine learning to address critical aspects of cancer research. These advancements have revolutionized our ability to analyze a large amount of biological data and also with unprecedented resolution, allowing us to analyze every single cell and detect the presence of even the smallest percentage of circulating tumor DNA fragments in liquid biopsies. Indeed, the increase in the accuracy and cost reduction of sequencing led to an exponential increase in the amount of data generated, which allows the creation of large-scale human genomics, transcriptomics, and proteomics datasets, allowing and requiring the need to develop novel computational methods capable of analyzing this large-scale data that can help advance our understanding of cancer biology, promising to improve diagnostic accuracy and therapeutic strategies for cancer patients.

The work presented in this thesis is divided into three main chapters: in the first chapter, a method is presented to segregate non-malignant tumor microenvironment (TME) cells from malignant ones and characterize tumor heterogeneity at high resolution by automatically identifying clonal copy number substructure from single cell RNA-seq data; In the second chapter, the potential of fragmentomics combined with deep learning models in early cancer diagnosis and minimal residual disease (MRD) analysis based on liquid biopsy data is explored; finally, the last chapter describes a method based on Gaussian processes for the prioritization of oncology drug targets based on a proteomic dataset.

Keywords: Liquid biopsy, Machine Learning, Early cancer detection, Minimal residual disease, Tumor heterogeneity, Drug Prioritization

Sintesi in lingua italiana

La tesi di dottorato presenta un quadro di analisi completo che sfrutta i progressi delle tecnologie di sequenziamento e dell'apprendimento statistico automatico per affrontare gli aspetti critici della ricerca sul cancro. Questi progressi hanno rivoluzionato la nostra capacità di analizzare una grande quantità di dati biologici e anche con una risoluzione senza precedenti, permettendoci di analizzare ogni singola cellula e di rilevare la presenza anche della più piccola percentuale di frammenti di DNA tumorale circolante nelle biopsie liquide. In effetti, l'aumento della precisione e la riduzione dei costi del sequenziamento hanno portato a un aumento esponenziale della quantità di dati generati, che consente la creazione di insiemi di dati genomici, trascrittomici e proteomici umani su larga scala, consentendo e richiedendo lo sviluppo di nuovi metodi computazionali in grado di analizzare questi dati su larga scala che possono contribuire a far progredire la nostra comprensione della biologia del cancro, promettendo di migliorare l'accuratezza diagnostica e le strategie terapeutiche per i pazienti affetti da cancro.

Il lavoro presentato in questa tesi è suddiviso in tre capitoli principali: nel primo capitolo viene presentato un metodo per segregare le cellule del microambiente tumorale (TME) non maligne da quelle maligne e caratterizzare l'eterogeneità del tumore ad alta risoluzione identificando automaticamente la sottostruttura del numero di copie clonali da dati RNA-seq di singole cellule; Nel secondo capitolo viene esplorato il potenziale della frammentomica combinato a modelli di deep learning nella diagnosi precoce del cancro e nell'analisi della malattia minima residua (MRD) sulla base di dati di biopsia liquida; infine, l'ultimo capitolo descrive un metodo basato sui processi gaussiani per la prioritizzazione di bersagli farmacologici oncologici sulla base di un set di dati proteomici.

Parole chiave: Biopsia liquida, Machine Learning, Individuazione

precoce del cancro, Malattia minima residua, Eterogeneità tumorale, Prioritizzazione dei farmaci

Acknowledgements

The research presented in this dissertation has received funding from AIRC under 5 per Mille 2018—ID. 21073 project—P.I. Maio Michele, G.L. Ceccarelli Michele. The research leading to these results has received funding from Italian Ministry of Research Grant PRIN 2017XJ38A4_004 and Associazione Italiana per la Ricerca sul Cancro (AIRC) IG grant 2018 project code 21846.

Contents

List of Acronyms

The following acronyms are used throughout the thesis.

ML Machine Learning

RQ Research Question

List of Figures

List of Tables

Introduction

The following chapters present computational algorithms that leverage the latest advancements in sequencing technologies and statistical machine learning to address critical aspects of cancer research. Advances obtained thanks to recent technologies have revolutionized our ability to analyze a large amount of biological data with unprecedented resolution. With the increase in the accuracy and cost reduction of sequencing, there has been an exponential increase in the amount of data generated, which has led to the creation of large-scale human genomics, transcriptomics, and proteomics datasets.

Innovations in sequencing have led to the creation of new computational techniques that are capable of analyzing large-scale data, which can help advance our understanding of cancer biology. They allow us, for example, to analyze every single cell and detect even the smallest percentage of circulating tumor DNA fragments in liquid biopsies.

The objective of the thesis is to develop novel computational algorithms based on machine learning to analyze this innovative data. The research aims to improve diagnostic accuracy and therapeutic strategies for cancer patients.

Specifically, in the first chapter, our algorithm for the automatic analysis of intratumor heterogeneity from single-cell transcriptomics data is described. The algorithm segregates malignant cells from non-malignant cells and then performs a comprehensive downstream analysis to automatically identify tumor subclones and classify distinct and shared alterations. The individual cells' smoothed expression profiles provide evidence of the copy number profile in each subclone, enabling accurate discrimination be-

tween malignant and non-malignant cells. This algorithm makes use of a multichannel segmentation technique that assumes that all cells in a given copy number clone have the same breakpoints. The copy number status of each segment is called by the mixture model algorithm, for which the parameters are estimated using the EM algorithm. The proposed algorithm has been successfully compared with respect to state-of-the-art methods to analyze tumors and their microenvironment using datasets from different tumor types and technologies.

Chapter two delves into the exciting realm of detecting cancer at an early stage and analyzing minimal residual disease (MRD) through liquid biopsy data. In this chapter, we explore the potential of using deep learning models based on fragmentomics features to achieve this goal. Fragmentomics features refer to the characteristics of DNA fragments that are present in a sample, and they can be used to identify subtle changes in the genome that may indicate the presence of cancer or MRD. By utilizing deep learning models, we can train the system to recognize patterns and make accurate predictions based on the fragmentomic features. This approach has the potential to greatly enhance the accuracy and speed of cancer detection, as well as MRD analysis, which is crucial for monitoring the success of cancer treatments and preventing relapse. The use of liquid biopsy data is particularly exciting, as it offers a minimally invasive alternative to traditional tissue biopsies. With liquid biopsies, a small sample of blood or other bodily fluids can be analyzed for the presence of cancer or MRD, making the process less invasive and more accessible to patients.

Finally, in the last chapter, we explore the prioritization of oncology drug targets, and this involves utilizing advanced techniques to identify and analyze the proteins that are expressed in cancer cells. By understanding the features of approved drug target proteins in cancer, we can prioritize and identify potential new drug target proteins according to their similarity with known approved, which could be used to develop more effective cancer treatments. Our research will involve analyzing large datasets of proteomics data and using machine learning approaches to identify patterns and relationships between different proteins. Our goal is to provide insights into the most promising targets for oncology drug development and to help accelerate the development of new and more effective cancer treatments.

Chapter 1

Intratumoral heterogeneity in scRNA

1.1 Single cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is a ground-breaking technology that, since its introduction in 2009 [96], has revolutionized the field of genomics. It enables researchers to study gene expression at the single-cell level, providing unprecedented insights into the biological processes that occur within individual cells, significantly influencing research in the fields of cancer biology and immunology. Unlike traditional bulk RNA sequencing methods, which provide an average gene expression profile for a population of cells, scRNA-seq allows for the identification of cellular heterogeneity (Figure 1.1), which in turn enables researchers to gain a more comprehensive understanding of the diversity of cell types and states within a tissue or tumor. Smart-seq2 [72] and 10X Genomics Chromium are the two most popular platforms used for scRNA-seq. Smart-seq2 [72] is based on microtiter plates, where mRNA is isolated and reverse transcribed to cDNA for each cell. In contrast, the 10X Chromium protocol uses a droplet-based scRNA-seq technology (3'-tag sequencing) that captures the gene expression profiles of thousands of cells simultaneously. In particular, they differ in terms of the number of captured cells, library sizes, and genes per cell. 10x protocol captures more cells, while Smart-seq2 [72] protocol captures more genes per cell.

By analyzing gene expression at the single-cell level, scRNA-seq has the potential to facilitate the discovery of new cell types and subtypes (Figure 1.1), as well as to help identify novel therapeutic targets for a wide range of diseases.

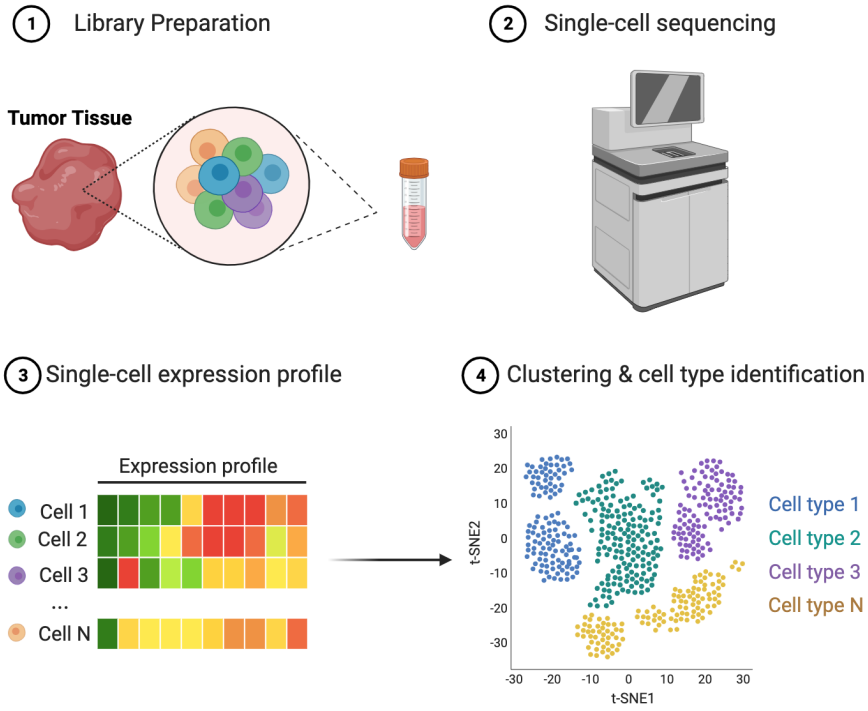


Figure 1.1. scRNA Workflow. Representative example of the single-cell sequencing analysis pipeline.

1.2 Intratumoral heterogeneity

The intratumoral heterogeneity and interactions between tumor cells and the immune system are a key step to explaining treatment failure and play a crucial role in studying tumor growth and evolution [1, 6], therefore, the analysis and study of these processes represent the main

opportunity to understand the reasons for treatment ineffectiveness and for the development of new personalized therapies.

scRNA-Seq has achieved remarkable success in delineating numerous transcriptional programs activated within a single tumor [94, 70, 28] and in identifying key regulators of tumor-host interaction. To study the complexity of lineage identity, differentiation, and proliferation of tumor cells and the impact of stromal and immune components, a large number of unsorted cells from tumor biopsies are subject to whole transcriptomics profiling and then classified as malignant cells, stromal cells, and immune cells, and further stratified into different compartments according to either expression of specific markers [11], and the orchestrated activation of pathways [28].

The primary stage typically is to classify cells into malignant cells, stromal cells, and immune cells and further stratified into distinct compartments based on the expression of specific markers and orchestrated activation of pathways. Differentiation between malignant and non-malignant cells is a critical step in the subsequent analysis of scRNA-seq tumor datasets.

1.3 Copy number inference

The fundamental approach to address the challenge described above is to estimate common copy number alterations that characterize transformed cells. Copy number alterations (CNAs) refer to a type of genetic mutation that involves changes in the number of copies of certain DNA segments within a cell. These mutations can occur in somatic cells, which are non-reproductive cells that make up most of the body and are commonly observed in various cancer types. CNAs can result in the loss or gain of genomic material, which can disrupt the normal function of genes and contribute to the development and progression of cancer. As such, understanding the mechanisms underlying CNAs and their role in cancer biology is critical for developing effective diagnostic and therapeutic strategies. Copy number alterations (CNAs) are somatic changes to chromosome structure that result in gain or loss in copies of sections of DNA and are prevalent in many types of cancer.

The copy number profile is inferred in single-cell transcriptomics by

considering the gene expression profiles of individual cells as a function of their genomic coordinates. The gene expression function is then subjected to moving average smoothing and grouped into malignant and non-malignant cells. One of the most successful methods based on this approach is the inferCNV algorithm [70]. One drawback is that the clusters of reference cells require manual identification, usually with a combination of approaches [66, 99]. Moreover, inferCNV and similar methods [25, 70] are particularly suited for smart-seq data having high coverage and relatively low throughput, whereas they exhibit sub-optimal performances on droplet-based methods with very sparse coverage depth and higher throughput [114]. An approach to overcome these limitations is represented by the CopyKAT method [27] that automatically classifies malignant and non-malignant cells. It was successfully applied to analyze the clonal substructure of three triple-negative breast tumors. However, the classification produced by CopyKAT can be affected by a wrong identification of normal cells and, similarly to other methods, was not designed to perform a complete automatic identification of the clones, reporting their breakpoints, the specific and shared alteration, and a clonal deconvolution in a complete end-to-end pipeline.

1.4 SCEVAN

The computational approach to the problem described in this thesis is called Single CEll Variational Aneuploidy aNalysis (SCEVAN), an algorithm based on variational principles for automatically identifying clonal copy number substructures within tumors from single-cell data. The method independently segments malignant cells from non-malignant cells and then analyses clusters of malignant cells using an optimization-based joint segmentation algorithm **??**. The concept that all cells within a copy number clone share identical breakpoints with the lysed expression profile of each cell has been exploited, providing support for defining the copy number profile of each subclone. Consequently, the joint segmentation process improves the rectification of systematic biases, resulting in consistent breakpoints. Subsequently, SCEVAN performs a comprehensive downstream analysis to automatically identify tumor subclones, classify distinct and shared alterations, and construct a phylogeny of clones. The joint segmen-

tation algorithm incorporated in SCEVAN has its roots in a variational framework originally developed in the field of Computer Vision, relying on the Mumford-Shah energy model [64]), which has already proven effective in identifying copy number alterations in matched tumor-normal pairs using high-density comparative genomic hybridization arrays [62] and in identifying fusion breakpoints [87]. Moreover, its joint version was developed to identify recurrent copy number alterations in large tumor cohorts [63, 61].

SCEVAN uses a set of stromal and immune signatures and the fact that malignant cells often harbor aneuploid copy number events to discriminate between transformed cells and microenvironment cells automatically. An extensive collection of annotated datasets of different tumor types was used, confirming that SCEVAN is faster and more accurate than state-of-the-art methods. The evaluation has shown that this approach is viable in cases with high purity and subjects with a significant amount of immune infiltration. Therefore, SCEVAN is particularly suited in studies where unsorted populations of single cells need to be analyzed to characterize, for example, the interaction between malignant cells and their microenvironment.

Subsequently, the key use of SCEVAN consists of delineating the clonal substructure in solid tumors based on differences in CNAs and studying tumors' temporal and geographic evolution. This includes the reporting of breakpoints, specific and shared alterations, and clonal deconvolution within a complete end-to-end pipeline. For instance, SCEVAN was used to deconvolve the clonal structure of glioma tumors, and in one patient, the presence of cell populations with differential activation of glioma cellular states was found, confirming that the clonal architectures drive the heterogeneity of glioma subtypes [28]. Functional analysis of subclones revealed drivers of cellular states, such as the Proliferative/Progenitor (PPR) glioma subtype. UBE2T was identified as the top amplified and differential expressed gene in the PPR clone. Interestingly, UBE2T can be pharmacologically inhibited [113], and therefore, it results as a potential therapeutic target for PPR cells. Moreover, SCEVAN can characterize the clonal status of onco-suppressor genes such as PTEN and CDKN2A. Such characterization may be of interest for diagnostics or therapeutic targeting and for the exploitation of approaches based on synthetic lethality[39].

Clonal deconvolution extracted from scRNA-seq can also be used to study regional and temporal tumor evolution and for the characterization of the difference between primary and metastases. SCEVAN has been evaluated with different single-cell technologies and recently used in a large study integrating millions of single cells from 538 samples and 309 patients across 29 datasets using the most commonly applied platforms such as 10x Chromium, Smart-seq2, GEXSCOPE, inDrop, and Drop-Seq [78]. SCEVAN was benchmarked against state-of-the-art methods and demonstrated its superior performance in terms of speed and accuracy on real-world data with reference copy number information from bulk tumor profiles, and also on synthetic data. Some limitations of SCEVAN rely mainly on its basic assumption that their aneuploidy can identify cancer cells.

1.4.1 Workflow

The workflow of SCEVAN, illustrated in Figure 1.2, is described here in summary, while each individual step of the algorithm will be explained in detail in subsequent sections. SCEVAN starts from the raw count matrix with genes on rows and cells on columns. The input count matrix is log-transformed and then pre-processed by removing cells with a low number of detected transcripts and selecting the most expressed genes. A set of highly confident non-malignant cells is identified and used to determine a copy number baseline and to compute the relative matrix removing the baseline (Steps A and B). This matrix undergoes an edge-preserving non-linear diffusion filter assuming a piece-wise smooth function as the underlying model (Step C). The smoothed matrix is then segmented using the joint segmentation algorithm to obtain a copy number matrix (Step D). SCEVAN discriminates the normal cells from tumor cells as those falling in the cluster containing the highest number of confident normal cells (Step E). The different subclones are obtained by analyzing the clusters of the tumor cells in the Copy Number Matrix (Step F). Then, each cluster is segmented independently from the smoothed matrix to obtain a copy number profile for any subclone (Step G). The segments are classified in one of five predefined copy number states: deletion, loss, neutral, gain, or amplification, using a majority vote applied to a mixture model classification of each cell. Finally, SCEVAN characterizes truncal, shared, and clone-specific alterations, comparing different clusters and performing

enrichment analysis up to a clone phylogeny (Step H).

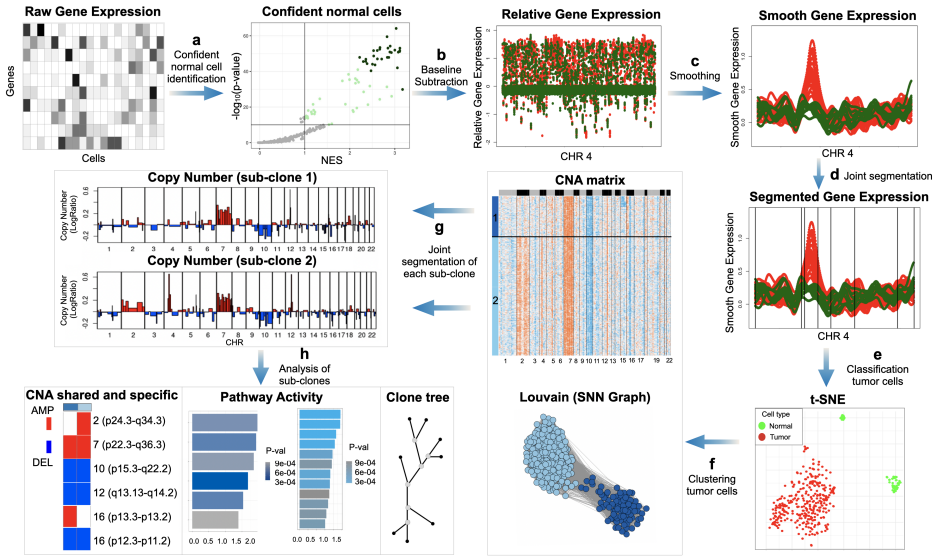


Figure 1.2. SCEVAN Workflow. SCEVAN starts from the raw count matrix, removing irrelevant genes and cells. **a** Identification of a small set of highly confident normal cells. **b** Relative gene expression obtained from removal of the baseline inferred from confident normal cells. **c** Edge-preserving non-linear diffusion filtering of relative gene expression. **d** Segmentation with a variational region growing algorithm. **e** Identification of normal cells as those in the cluster containing the majority of confident normal cells. **f** Identification of possible subclones using Louvain clustering applied to a shared nearest-neighbor graph of the tumor cells. **g** Segmentation with a variational region growing algorithm applied to each subclone. Segments are then classified into five copy number states. **h** Analysis of subclones including clone tree, pathway activities (GSEA was performed for each subclone using fgsea-Multilevel, which calculates p-values based on an adaptive multilevel splitting Monte Carlo scheme), and characterization of shared and specific alterations.

1.4.2 Preprocessing of scRNA-seq data

The preprocessing phase is aimed at filtering out low-quality and irrelevant cells. Specifically, the cells with less than 200 detected genes and

the genes expressed in less than 1% of cells are removed. The remaining genes are annotated by adding their genomic locations to the matrix using Ensembl-based annotation package [76], and then genes are sorted according to genomic coordinates. After annotation, the genes involved in the cell cycle pathway, obtained from REACTOME [37], are filtered to reduce artificial segments caused by the cell cycle [27].

1.4.3 Identification of High confident non-malignant cells

The input data D is an $m \times n$ single-cell gene expression matrix where m is the number of cells, and n is the number of genes ordered by genomic positions. To segregate malignant from non-malignant cells, SCEVAN follows a multi-step approach. A small set of high-confidence normal cells is used to build a relative expression matrix and as a seed for identifying the cluster of normal cells. Then, the relative expression matrix is segmented and clustered as described in the following paragraphs. A set of gene signatures from public collections [11, 111], including cells of the tumor microenvironment, stromal and immune cells, such as lymphocytes, macrophages, microglial cells, dendritic cells, neurons, and others, is , is used to identify the high-confidence normal cells. The Mann-Whitney-Wilcoxon single sample gene set test implemented in the yaGST package [26] was applied and assumed as normal confident cells the top classified cells with p-value less than 10^{-10} and Normalized Enrichment Score (NES) greater than 1.0. The search is restricted to a maximum of 30 high-confidence non-malignant cells.

Then the copy number baseline, estimated from the median expression of confident normal cells, is removed from the count matrix, thus obtaining the relative matrix $D_r = D - \tilde{\mathbf{b}}^T$ where $\tilde{\mathbf{b}}$ is the n-dimensional vector with the median value of confident normal cells. If no confident normal cells are found, is assumed that the sample is pure and contains only malignant cells. In this case, a synthetic baseline is removed from the malignant cells. The synthetic baseline is obtained by subtracting from each gene a random value extracted from a Gaussian distribution with zero mean and the same standard deviation of the considered gene. To take into account the heterogeneity of the sample and to avoid smoothing CNV subclones, this step is applied to clusters of the count matrix. The number of clusters is automatically chosen by using the Calinski-Harabasz criterion

and hierarchical clustering based on Ward's method, they exhibited better results in the analyzed data.

From now on, the relative gene expression matrix will be considered the sampled version of a function u defined on the genome with values in $\mathbb{R}^m$. In the case of single-cell data, the sampling is based on the relative expression values of each gene, in previous works, a similar formalism was used for aCGH arrays [61] where the sampling points are the position of each SNP probe, or for Whole Exome data [2] the sampling points are the genomic positions position of exons.

1.4.4 Edge-preserving smoothing

Before the segmentation phase, one of the key steps of SCEVAN is to smooth the relative expression function. Since the segmentation step described below assumes a piecewise-constant model of the copy number signal, a non-linear smoothing of the gene expression is preliminarily performed along with the genomic coordinates to regularize the gene expression signal, reduce the outliers and, at the same time, to preserve the discontinuities which are the breakpoints between the copy number segments. A filter grounded in the Bayesian framework of edge-preserving regularization was used, [17], which considers the minimization of the Total Variation (TV) functional

$$\int \phi(|\nabla u|) \tag{1.1}$$

where u is the m-dimensional relative gene expression signal, ∇u is its gradient and $\phi(\cdot)$ is a discontinuity-adaptive prior [12]. In particular, a TV approximation with a parametric family of functionals that are smooth at the origin is used:

$$\kappa \int \log\cosh(\frac{|\nabla u|}{\kappa}) \tag{1.2}$$

which has been shown to produce a well-posed minimization problem overcoming the non-differentiability of the TV at the origin [14], where $\kappa > 0$ is the regularization parameter that measures the amount of smoothing performed by the algorithm, which increases as the value approaches 1.

The discretization of the minimization solution of TV is based on the use of a finite difference method for smoothing data while preserving edges. The iterative numerical scheme implemented in SCEVAN uses a hyperbolic conservation law-based technique that is just the one-dimensional adaptation of the stable finite difference scheme previously reported [14]. Specifically, once defined $\Delta^x_{\pm} = \pm(u_{i\pm1,j} - u_i)$ and Δt in this case is a fixed value that represents the time step, then in each subsequent iteration of approximation $n+1$, the smoothed relative gene expression signal is computed as follows:

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x} \cdot \Delta^x_- \left\{ \tanh \frac{\Delta^x_+ u_i^n}{\kappa \Delta x} \right\} \qquad i = 1, \ldots, m \tag{1.3}$$

1.4.5 Single Cell joint segmentation algorithm

SCEVAN uses a multichannel segmentation procedure that inputs all the cells in a given clone to identify the boundaries of homogeneous copy numbers.

The procedure is based on the *Mumford and Shah energy* originally developed to analyze images. In their original work [64], the authors introduced the basic properties of variational models for computer vision aimed at defining the mathematical foundations for appropriate decomposition of the 2D domain Ω of a vector-valued function $\mathbf{u_0} : \Omega \to \mathbb{R}^m$ into a set of disjoint connected components ($\Omega = \cup_{i=1}^{l} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \quad 1 \leq i, j \leq l, i \neq j$). The set of points on the boundary between the Ω_i is denoted as Γ. This partition is modeled such that the signal varies smoothly within a component and discontinuously between the disjoint components. This problem is known as piece-wise smooth approximation. Here a special case of the Mumford-Shah model was adopted, when the approximation $\mathbf{u}$ of the signal $\mathbf{u_0}$ is constrained to be a piece-wise constant function.This is best suited for CNV segmentation. In this case, the optimal segmentation is obtained by minimizing the following:

$$E(u, \Gamma) = \sum_i \int_{\Omega_i} (\mathbf{u_0} - \mathbf{u_i})^2 \mathrm{d}x\mathrm{d}y + \lambda |\Gamma| \tag{1.4}$$

where Γ is the boundary between the connected components Ω_i and $|\cdot|$

indicates its length and $\mathbf{u_i}$ is the restriction of $\mathbf{u}$ to Ω_i. It is easy to show that the minimum for this model can be obtained by posing $\mathbf{u_i}$ as the mean of $\mathbf{u_0}$ within each connected component Ω_i. Hence, this functional represents a compromise between the accuracy of the approximation and the parsimony of the boundaries. It is essential to notice that the resulting segmentation depends on the scale parameter λ. Indeed, it determines the number of computed regions: when λ is small many boundaries are allowed, and the resulting segmentation will be fine. As λ increases, the segmentation will be coarser and coarser.

In the current case of segmenting the genome in regions of homogeneous copy number, a segmentation $\Gamma = \{b_1, \cdots, b_{M+1}\}$ was defined as a set of ordered positions (breakpoints) partitioning the genome into M connected regions $R = \{R_1, \cdots, R_M\}$. Each region R_i will contain all genes whose genomic coordinates lie between breakpoints $\{b_i, b_{i+1}\}$. A function defined on a one-dimensional domain in equation (1.4) was modeled, $|\Gamma|$ reduces to the number of regions M. According to the original algorithm proposed in [61], to minimize this function, adjacent regions R_i and R_{i+1} are iteratively merged in a pyramidal manner to create larger segments, and the reduction of the energy can be shown as:

$$E(u, \Gamma \backslash \{b_i\}) - E(u, \Gamma) = \frac{|R_i||R_{i+1}|}{|R_i| + |R_{i+1}|} ||\mathbf{u_i} - \mathbf{u_{i+1}}||^2 - \lambda \tag{1.5}$$

where $|R_i|$ is the length of the i-th region, and $\mathbf{u_i}$ is a m-dimensional vector with the mean value of gene between b_i and b_{i+1}, $||\cdot||$ is the L_2 norm and $\backslash$ is the set difference. To minimize (1.4), a greedy procedure was followed. A segmentation is started with n regions, one for each gene. Then, at each step, the adjacent regions that yield the maximum decrease of the energy functional upon merging are merged. Since λ decides the end of merging, choosing an appropriate value is crucial to ensure the quality of the final segmentation. As in [61], the selection for λ at each merging step is done dynamically, depending on two factors: the region's size and the mean values of the consecutive regions being considered for the merge. Hence, the cost of merging two regions R_i and R_{i+1}, associated with a breakpoint b_i, is computed as follows:

$$\tilde{\lambda}_i = \frac{|R_i|\,|R_{i+1}|}{|R_i| + |R_{i+1}|}||\mathbf{u_i} - \mathbf{u_{i+1}}||^2, \tag{1.6}$$

if $\tilde{\lambda}_i < \lambda$, the adjacent regions are merged and the i-th breakpoint removed. Otherwise, the energy function has reached a local minimum, and no merging can be done further. Therefore, λ is updated to the smallest of $\lambda_i + \epsilon$, continuing the merging. The sequence of λ values is monotonically increasing as it corresponds to the amount of decrease of the energy functional at each step in (eq. (1.5)). In [62] a stopping criterion was adopted in such a way that the final segmentation is obtained when the increase in λ stabilizes and merging any further does not correspond to a significant decrease of the energy. The final stopping value is based on the variability of the adjacent region and the total variability of the data, ν. The total variability is computed as the sum of the standard deviation of all cells after the smoothing step. The stopping criterion is $\Delta\lambda = \lambda_{i+1} - \lambda_i \leq \beta\nu$, where β is a positive constant, representing the only parameter of the segmentation algorithm.

1.4.6 Classification of malignant and non-malignant cells

The joint segmentation algorithm, applied to the relative gene expression matrix, returns a set of breakpoints and the interpolating function u minimizing (1.4), which is simply the mean gene expression between consecutive breakpoints in each cell. Hence, an intermediate CNA $m \times n$ matrix(m is the number of cells and n is the number of genes) is computed by substituting each expression value with the mean gene expression between consecutive breakpoints in each cell. This matrix is then clustered into two groups using hierarchical clustering. All the cells in the cluster containing the highest number of confident normal cells (if confident normal cells have been detected as described above) are then classified as non-malignant. The final CNA matrix is then obtained by subtracting the vector of the mean value of all the identified normal cells.

1.4.7 Differential subclonal structure characterization

To deconvolve the clonal structure of a given sample, the CNA matrix containing just tumor cells is clustered using Louvain clustering [10] ap-

plied to a shared nearest-neighbor graph [109] (Figure 1.2, step F). Clonal deconvolution is performed with a standard technique for identifying subpopulations in the analysis of single-cell transcriptomic data. Each cluster represents a potential subclone. Therefore the joint segmentation algorithm is re-applied considering just the cells of the cluster (Figure 1, step G). The segmentation results are classified with the CNV calling algorithm described below and analyzed to identify subclone-specific alterations, shared alterations between subsets of clones, and clonal alterations. Segments in each clone representing the same copy number alterations at genomic distances less $10Mb$ are first merged together. Afterwards, two alterations in different clones are considered the same if the respective start or end breakpoints are at a genomic distance of less than $10Mb$ and differ in size by less than 40%. This choice prevents altered segments that show small differences in the segmentation, as close but not equal breakpoints, are considered as different alterations. The parameters mentioned were selected based on a grid search, selecting the parameters that showed better results from the analysis of downstream results of all samples analyzed. Finally, the list of potential clone alterations is further filtered, retaining only clones having specific alterations.

1.4.8 CNV calling

To obtain an estimate of the copy number status of each segmented region, a mixture model-based algorithm to the mean expression level $u_{i,j}$ from CNA matrix of each $i-th$ cell within each $j-th$ segment was applied. This value is modeled as a mixture of five truncated normal distributions (with lower bound a and upper bound b) as in [58], each defined as follows:

$$\mathscr{T}_a^b\left(u_{i,j};\theta\right) = \frac{f\left(u_{i,j};\theta\right)}{F(a;\theta) - F(b;\theta)} I_a^b\left(u_{i,j}\right) \tag{1.7}$$

where $I_a^b(\cdot)$ represents the indicator function of belonging of $u_{i,j}$ to the interval, $f(\cdot)$ density and $F(\cdot)$ cumulative distribution functions.

The parameters of the mixture are estimated using the EM algorithm [21], starting at step $t = 0$ from empirically chosen initial fixed parameters (Table 1.1).

In the expectation step, the conditional probability $\xi_{s,i,j}^{t+1}$ is calculated

for each mixture (s), with actual parameters the prior p_s and $\theta_s = (\mu_s, \sigma_s^2)$.

$$\xi_{s,i,j}^{t+1} = \frac{p_s^{(t)} \mathscr{T}_a^b \left(u_{i,j}; \theta_s^t\right)}{\sum_{i=1}^{m} \sum_{j=1}^{r} p_s^{(t)} \mathscr{T}_a^b \left(u_{i,j}; \theta_s^t\right)} \tag{1.8}$$

where m is the number of cells and r the number of segments.

In the maximization step, defined the sum of all conditional probabilities as $\Xi_s^{t+1} = \sum_{i=1}^{m} \sum_{j=1}^{r} \xi_{s,i,j}^{t+1}$, the new parameters for each mixture are estimated as follows:

$$\mu_s^{t+1} = \frac{\Xi_s^{t+1} u_{i,j}}{\Xi_s^{t+1}}, \quad \sigma_s^{t+1} = \sqrt{\frac{\Xi_s^{t+1} \left(u_{i,j} - \mu_s^{t+1}\right)^2}{\Xi_s^{t+1}}}, \quad p_s^{t+1} = \frac{\Xi_s^{t+1}}{m * r}$$

Then, each segmented region is classified, using a posterior probability, in one of five copy number states: deletion (0), loss (1), neutral(2), gain (3), or amplification (4). The final classification of each segmented region is obtained using the majority vote algorithm, starting from the classification for each cell in the relative segment.

1.4.9 Comparison with other methods and analysis of bulk data

The raw count matrices of scRNA-seq samples reported in classification and copy number inference comparisons reported in the paper are analyzed following the steps of SCEVAN Workflow and with CopyKAT v1.0.5 and inferCNV v1.4.0. InferCNV was run using the author's recommendations for the parameters `denoise=TRUE, HMM=TRUE, HMM_type='i6'`, and `cutoff=0.1` (for MM dataset)[54], `cutoff=1.0` (for multiregional GBM dataset)[70].

The copy number variation profile from bulk biopsies was used as ground truth. In the case of Multiple Myeloma [54], CNVkit v0.9.9 was used for segmentation. The integer Copy Number was assigned based on cutoffs specified in the CNVkit documentation (-1, -0.25, 0.2, and 0.7).For the 26 Glioblastoma multiregional samples of low-depth whole-genome sequencing (WGS) on the bulk biopsies [112], the copy number variations computed every 1Mb window by Yu et al. [112] was segmented using DNAcopy (v1.62.0)[102].The ground truth extracted from WES and WGS have, of course, different resolutions with respect to the single-cell data. Therefore, the output of each method and the ground truth were

	MIXTURE 0 (DELETION)	MIXTURE 1 (LOSS)	MIXTURE 2 (NEUTRAL)	MIXTURE 3 (GAIN)	MIXTURE 4 (AMPLIFICATION)
BOUND (Clonal Analysis)	$[\mu_c[1]*3, \mu_c[1] - (3*\sigma)]$	$[\mu_c[1] - (3*\sigma), -\beta]$	$[-\beta, \beta]$	$[\beta, \mu_c[3] + (3*\sigma)]$	$[\mu_c[3] + (3*\sigma), \mu_c[3] * 3]$
BOUND (Subclonal Analysis)	$[\mu_{sc}[1]*3, \mu_s[1] - (3*\sigma)]$	$[\mu_s[1] - (3*\sigma), -\beta*2]$	$[-\beta*2, \beta*2]$	$[-\beta*2, \mu_{sc}[3] + (3*\sigma)]$	$[\mu_{sc}[3] + (3*\sigma), \mu_{sc}[3] * 3]$
μ_c (Clonal Analysis)	$-\beta*4$	$-\beta*2$	0.0	$\beta*2$	$\beta*4$
μ_{sc} (Subclonal Analysis)	$-\beta*4$	$-\beta*3$	0.0	$\beta*3$	$\beta*4$
σ	0.01	0.01	0.01	0.01	0.01
β	0.05	0.05	0.05	0.05	0.05
p	0.05	0.1	0.7	0.1	0.05

Table 1.1. Initial parameters of the mixture of five truncated normal distributions for copy number classification.

first resampled at the same genomic resolution. Specifically, for each position of the genome at 1Mb distance is taken the log ratio value or copy number integer value depending on the considered method. Then, the Pearson correlation is computed between this re-sampled vector and the ground truth [27]. CNVkit and DNAcopy use circular binary segmentation (CBS), which is not used by any of the methods SCEVAN, CopyKAT, and inferCNV compared. This choice avoids a possible bias in the comparison.

For the comparison of breakpoints detection on synthetic data, is used GenoCN v1.40.0 and the method doGFLars of jointseg v1.0.2. Since they do not have their own smoothing method, is used smooth.CNA of DNAcopy [102] as previously suggested [73].

1.5 SCEVAN benchmark

1.5.1 Malignant cell classification on synthetic data

To quantitatively evaluate the accuracy of SCEVAN in discriminating malignant from non-malignant cells, 500 synthetic matrices with known tumor/normal classification were generated. A Multiple Myeloma dataset has been used, containing 17,267 malignant plasma cells and 57,719 immune cells of Liu et al. [54]. Based on the specific markers used by the authors, cell clusters are classified into eight immune compartments and tumor cells of each patient. A scDesign2[91] model for each cell type was trained, specifically eight immune and 14 malignant models, one for each sample. The synthetic scRNA-seq matrices were randomly generated by choosing the following parameters: the number of total cells (between 300 and 1000), the tumor purity (between 5% and 100%), the number of cells for each immune cell type, and the scDesign2[91] malignant model from one of the 14 samples. The generated matrices had, on average, 94% of zero

values. Dropout noise at different levels was also added to each simulated sparse count matrix. Dropout simulations have probabilities conditioned on mean gene expression, such that lowly expressed genes have a higher likelihood of dropout than highly expressed genes. This type of noise is added using SPLATTER[115], which uses a logistic function to produce a probability that a count should be zero. The logistic function is defined by a midpoint parameter, x_0, the logarithm of the expression level at which 50% of cells are replaced with zero. The probability of a zero for each gene is then used to randomly replace some of the simulated counts with zeros using a Bernoulli distribution. Three noise levels have been used, corresponding to the values of $x_0 = -2, -1, 0$, that respectively replace 7%, 17%, and 31% of non-null values with a 0. SCEVAN and CopyKAT have been applied to these synthetic matrices containing a total of 322,687 cells (Figure 1.3), obtaining with SCEVAN a mean F1 score of 0.948 - 0.943 - 0.909 - 0.824 and with CopyKAT 0.798 - 0.792 - 0.763 - 0.726 respectively for each level of noise. It is worth noticing that in some cases, both methods can obtain a very low F1 score, this is due to the fact that in cases of erroneous identification of the cluster of normal cells, for example, a cluster of tumor cells is named as the reference normal, then a complete mis-classification can happen and an F1 score close to zero is obtained.

1.5.2 Malignant cell classification accuracy on real data

The accuracy of non-malignant cell classification also has been evaluated on real data, the method was applied to several public datasets [66, 114, 49, 75, 112] of three different cancer types of scRNA-seq data (Glioblastoma (GBM), Head and Neck Squamous Cell Carcinomas (HNSCC), Colorectal cancer) and from different sequencing technologies (Smart-seq2, 10X Chromium), classifying a total of 106 samples and $93,322$ cells. In all the considered datasets, the identification of the non-malignant cell has been reported by the authors through manual curation based on a combination of approaches using copy number profile [70], clustering, and cell markers. The results in terms of F1 score have been compared[74] with those obtained by using CopyKAT [27]. SCEVAN, as shown in Figure 1.4, achieves a better classification score in 63% of the samples, whereas CopyKAT performs better than SCEVAN in 23% of the samples. The F1 score for all samples obtained with SCEVAN is 0.90 in contrast to the F1

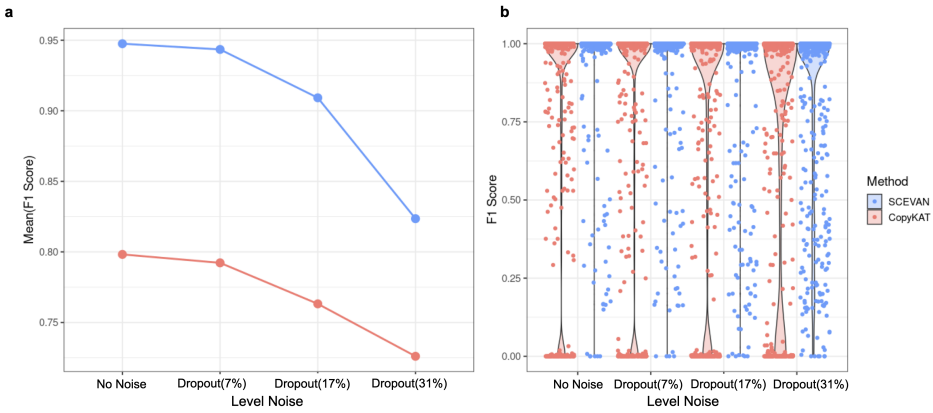


Figure 1.3. Benchmark of malignant cell classification task on synthetic data. **a** Comparison tumor/normal classification in terms of mean F1 score on 500 synthetic matrices, for various dropout noise levels, between SCEVAN and CopyKAT. (SCEVAN 0.948 0.943 0.909 0.824 - CopyKAT 0.798 0.792 0.763 0.726). **b** Violin Plots of the F1 score for 500 matrices at different noise levels using CopyKAT (Red) and SCEVAN (Blue).

score of 0.63 obtained with CopyKAT. SCEVAN shows a low F1 SCORE in samples with very few tumor cells (between 1 and 15), present mostly in the case of Head & Neck cancer dataset. For one of the samples (BT786), results could not be obtained from CopyKAT due to a crash. Collectively, these results confirm that SCEVAN can accurately discriminate between tumor and normal cells in different solid tumors using the copy number profiles inferred from scRNA-seq.

1.5.3 Segmentation accuracy on synthetic data

To perform a quantitative evaluation of the segmentation results, a synthetic dataset was generated that models two realistic scenarios: Scenario I, with just clonal alterations and all malignant cells share the same alterations; Scenario II, where there are some clonal alterations shared by all cells and also two populations of malignant cells having subclone-specific alterations. For both scenarios, synthetic matrices were generated with different levels of magnitude of the synthetic copy number alterations,

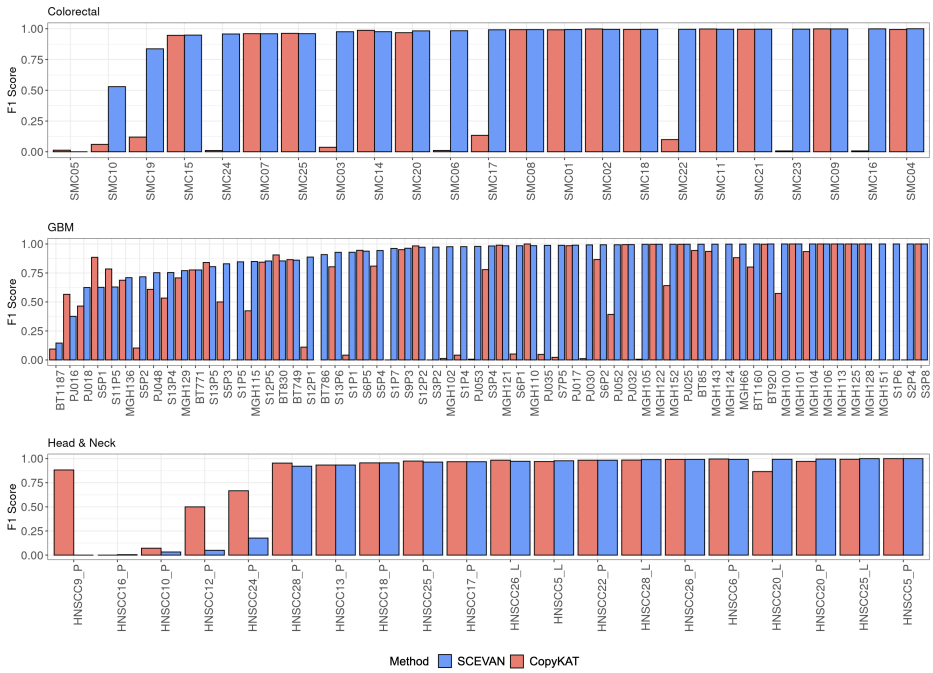


Figure 1.4. Benchmark of malignant cell classification task. F1 score obtained with SCEVAN and CopyKAT [27] in the classification of malignant and non-malignant cells for each cancer type. Colorectal cancer [49] n = 47,285 cells examined over 23 scRNA-seq independent experiments, Glioblastoma [66, 114, 112] n = 40,320 cells examined over 63 scRNA-seq independent experiments, Head and Neck Squamous Cell Carcinomas [75] n = 5,717 cells examined over 20 scRNA-seq independent experiments

starting from matrices previously obtained using scDesign2[91]. Only normal diploid cells are considered and randomly altered genomic regions, generating synthetic aneuploid cells.

For each matrix, the number of aneuploid cells was chosen randomly (between 30% and 90% of total cells), the number of alterations (between 1 and 10), the central position of each alteration (between 1 and the number of total genes), the number of genes belonging to each alteration (between 50 and 1000), and in the case of scenario II the assignment of each cell to one of the two sub-clones.

To generate synthetic amplification (deletion), the count values of the genes belonging to the alteration are increased (decreased). Specifically, a uniform random value ρ in $(0, \alpha)$ was drawn and replace each gene count x_{ij} by $x_{ij}(1+\rho)$ for amplifications and $x_{ij}/(1+\rho)$ for deletions. Therefore, are increased/decreased counts of the genes belonging to the alteration by a percentage between 0 and $100\alpha\%$. For each scenario, four experiments are performed corresponding to $\alpha = 2, 3, 4$, generating for each scenario, and value of α, 100 matrices.
To define an appropriate evaluation metric for the segmentation produced by various segmentation algorithms, as previously suggested [73], the breakpoints that lie within a tolerance threshold of distance (es. 20 genes) from the true breakpoints are considered as True Positive (TP), and as false negative (FN) if there are no breakpoints in this tolerance area. The synthetic dataset was used to compare the accuracy of SCEVAN and CopyKAT, and also other segmentation approaches were considered, such as GFLars [73], a method optimizing a squared loss and a regularization term based on group LASSO, and GenoCN, [92] a method based on HMM segmentation. Using a threshold of 20 genes, SCEVAN obtains significantly higher F1 scores than other methods in each scenario and experiment (Figure 1.5).

It is interesting to note that in some cases SCEVAN, as well as the other methods, gets a low score. This is due to several factors. When all the breakpoints are identified at a distance greater than the tolerance threshold, or the method fails to identify most of the alterations, then the corresponding classification score is close to zero. Moreover, since the synthetic matrices, as well as the synthetic alterations, are randomly generated, it is possible that the alterations are located in regions where the average gene expression is low. In such cases, even for the high amplitude of the alteration (the parameter α), the segmentation task becomes extremely challenging with the possibility of low detection accuracy.
The role of the parameters on the performance of the considered segmentation methods also needs to be investigated. In general, segmentation algorithms adopt some regularization parameters to control the amount of smoothing and the coarseness of the segmentation, such as the parameter β for SCEVAN that controls the convergence of the hierarchical region-merging procedure and defines a stopping criterion for the increasing sequence of the regularization parameters and KS.cut for CopyKAT.

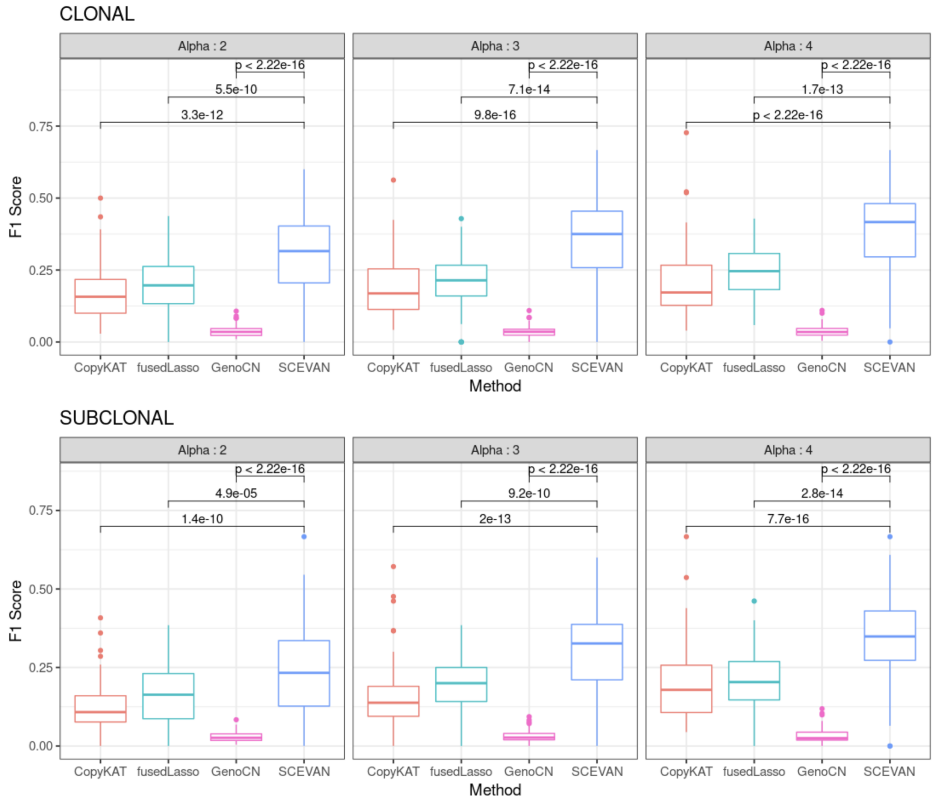


Figure 1.5. Benchmark of segmentation task on synthetic data. Comparison of segmentation accuracy in terms of F1 score with fixed tolerance threshold (20) as a function of the magnitude of alteration α, and type of alteration, Clonal (Scenario I) and Sublconal (Scenario II). Box plots show the median as the center, the lower and upper hinges that correspond to the 25th and the 75th percentile, and whiskers that extend to the smallest and largest value of no more than 1,5*IQR. Values that stray more than 1.5*IQR upwards or downwards from the whiskers are considered potential outliers and represented with dots. Significance was computed by a one-sided Wilcoxon signed rank test.

Since an exhaustive exploration of the parameters for the considered algorithms may lead to over-optimistic results which are difficult to replicate

in scenarios with real data, a dynamic programming approach was used that progressively selects optimal subsets of the breakpoints reported by a given method [73] (jpruneByDP procedure of the jointseg Bioconductor package). With this setting, it is possible to compute a precision-recall (PR) curve for the output of various algorithms varying the size of selected optimal subsets of breakpoints. Here, the mean area under the PR curve (AUC) was computed as a function of the tolerance parameter for 100 simulated matrices at different levels of the magnitude of alteration α (Figure 1.6).

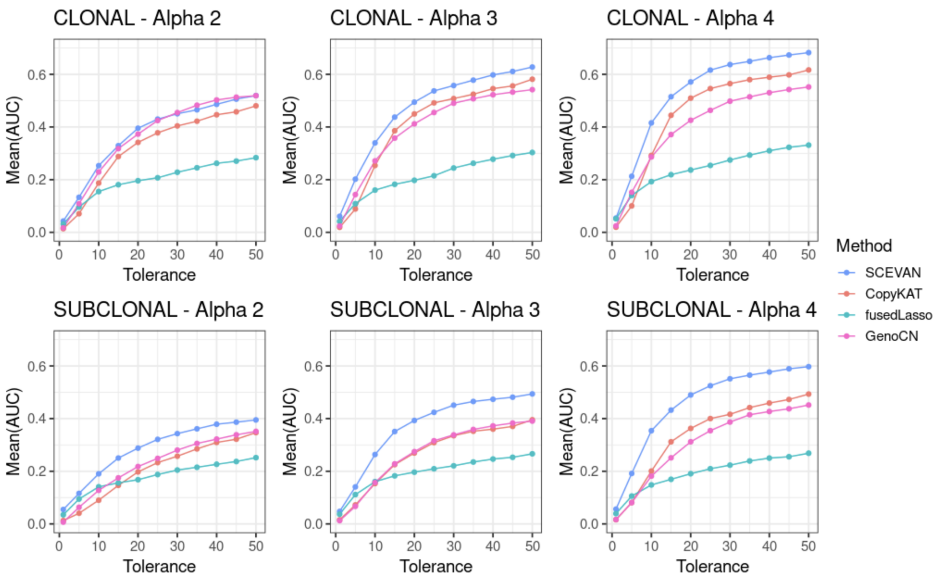


Figure 1.6. Segmentation accuracy on synthetic data. Comparison of segmentation accuracy in terms of mean PR AUC at different thresholds of tolerance on 100 synthetic matrices, for various magnitudes of alteration α, and type of alterations, Clonal (Scenario I) and Sublconal (Scenario II).

SCEVAN reaches consistently better AUC than the other segmentation methods, and as the α parameter increases, i.e., when the steps in the genomics alterations are more noticeable, the improvement is even more evident.
Furthermore, the performance varying the segmentation parameters of

SCEVAN and CopyKAT was evaluated. For CopyKAT, the parameter KS.cut has been varied in the interval suggested by the authors (0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.4), and for SCEVAN the parameter β (0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0) has been varied. In both cases, the increase of these values results in coarser segmentations. PR curves are calculated for matrices with different α (2, 3, and 4), with clonal and subclonal scenarios, and using different tolerance values (10, 20, 30, and 40 genes). This analysis also confirms that SCEVAN's accuracy is higher even with varying parameters and tolerance values (Figure 1.7). The results above refer to a limited number of alterations (between 1 and 10), it has been observed that have observed that the overall accuracy is not significantly influenced by the number of simulated genomic alterations. Rather, it is influenced by the magnitude of the alteration α and the local distribution of the smoothed gene expression signal around the discontinuities induced by the breakpoints. In the experiments reported in the sequel, the default value ($\beta = 0.5$) was used to produce slightly finer segmentations on real data accounting for more focal lesions. For the clonal analysis, the algorithm uses a slightly larger value ($\beta = 3.0$) to reduce the effect of the noise in the final output. Finally, the synthetic dataset is publicly available and could serve as a reference benchmark for other single-cell CNV inference algorithms.

1.5.4 Segmentation accuracy using reference data

After evaluating the accuracy of the method in the identification of the copy number breakpoints on synthetic data, Its accuracy on real datasets was evaluated. For this task, both the single-cell RNA-seq and reference copy number profiles obtained from bulk DNA sequencing were used. Since, in this case, real single-cell datasets were used, results produced by SCEVAN, inferCNV, and CopyKAT are compared. Since CopyKAT returns just the segment mean, whereas the output of inferCNV is the inferred copy number copy number status, when comparing both methods with SCEVAN both the segment mean, mentioned hereafter as LogRatio, and the copy number status called by the mixture model algorithm were used. The ground truth used is composed of 26 samples of a Glioblastoma multiregional dataset [112] with the CNV status from low-depth whole-genome sequencing (WGS) on the bulk biopsies (Figure 1.8c) and seven

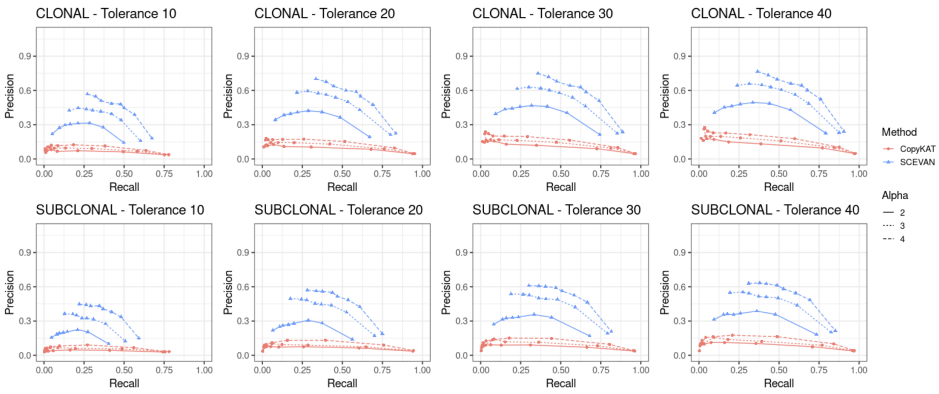


Figure 1.7. Precision-Recall of segmentation task on synthetic data at varying tolerance. PR curves obtained by varying the parameter β for SCEVAN and *KS.cut* for CopyKAT in terms of the magnitude of alteration α, level of tolerance (10,20,30,40) and type of alterations, Clonal (Scenario I) and Sublconal (Scenario II).

samples (81012 Primary, 59114 Relapse-1, 58408 Primary, 58408 SMM, 27522 Primary, 57075 Relapse-1, 37692 Primary) of Multiple Myeloma (MM) dataset [112] with the CNV Status obtained using whole-exome sequencing (WES) on the bulk biopsies (Figure 1.8d). The output of SCEVAN, CopyKAT, and inferCNV were re-sampled to the same resolution of the ground truth by taking one value every 1 Mb.

The boxplots of Figure 1.8 show the Pearson correlation between the inferred copy number profiles and the reference copy number obtained in all samples. SCEVAN as segment mean (LogRatio) has a mean correlation of 0.57 (max 0.81) on the multiregional GBM dataset and 0.44 (max 0.71) on the MM dataset. The copy number call of SCEVAN has a mean correlation of 0.54 (max 0.84) on the multiregional GBM dataset and 0.46 (max 0.76) on the MM dataset. CopyKAT has a mean correlation of -0.03 (max 0.52) on the multiregional GBM dataset and 0.29 (max 0.52) on the MM dataset. Whereas inferCNV has a mean correlation of 0.44 (max 0.88) on the multiregional GBM dataset and 0.35 (max 0.63) on the MM dataset.

Since inferCNV does allow automatic identification of the non-malignant cells, for the generation of these results, the set of non-malignant cells clas-

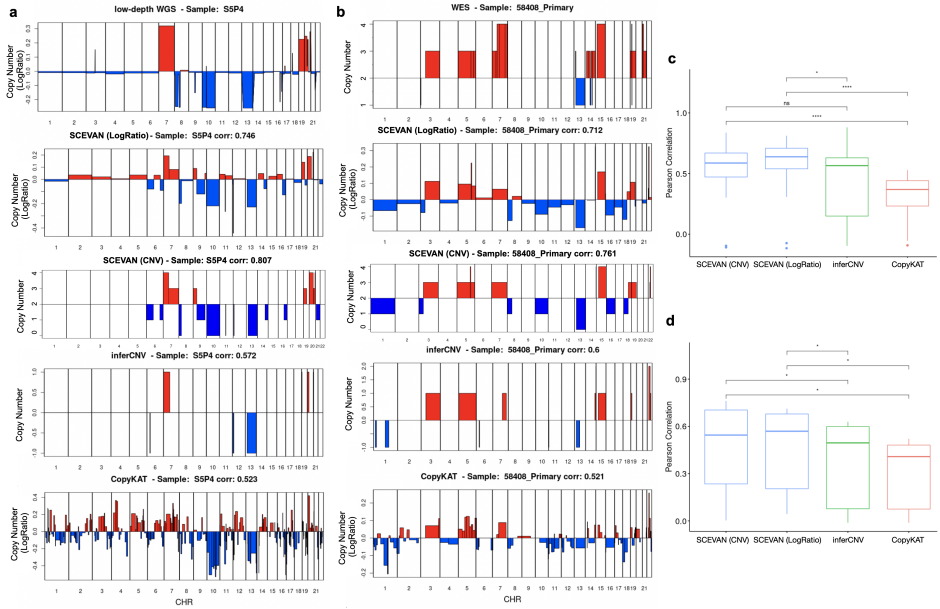


Figure 1.8. Benchmark of inferred copy number profile. (a,b) Copy number profile inferred with SCEVAN (segment mean (LogRatio) and CNV status), inferCNV, CopyKAT, the corresponding ground truth from low-depth WGS of sample S5P4 [112] and from WES of sample 58408 Primary [54]. In both **c** and **d** Box plots show the median as center, the lower and upper hinges that correspond to the 25th and the 75th percentile, and whiskers that extend to the smallest and largest value no more than 1,5*IQR. Values that stray more than 1.5*IQR upwards or downwards from the whiskers are considered potential outliers and represented with dots. Significance was computed by a two-sided Wilcoxon signed rank test (ns: p-value $>$ 0.05, *: p-value $<=$ 0.05, ****: p-value $<=$ 0.0001). **c** Pearson correlation between the copy number inferred with different methods and the ground truth from low-depth WGS for 26 samples [112]. SCEVAN obtains a significantly higher correlation than CopyKAT (LogRatio p-value $1.3e-05$ and CNV status p-value $3.0e-07$) and inferCNV (LogRatio p-value 0.02). **d** Pearson correlation with the ground truth from WES for 7 samples [54]. SCEVAN obtains a significantly higher correlation than CopyKAT (LogRatio and CNV status p-value 0.016) and inferCNV (LogRatio p-value 0.016 and CNV status p-value 0.031). Source data are provided as a Source Data file.

sified by SCEVAN was used. The lower accuracy of CopyKAT is probably due to the wrong classification of malignant and non-malignant cells. However, since the misclassification of normal cells could be eventually corrected by manual inspection, instead of using the whole multiregional dataset, [112] the same comparison has been performed using just the samples where CopyKAT achieves an F1 classification score above 0.50. This comparison evaluated the accuracy of segmentation on real-world data, limiting the effect of malignant/non-malignant misclassification. On the 13 samples where CopyKAT reaches the best classification results, SCEVAN obtained a median correlation between the inferred CNV profile and the CNV from the bulk WGS of 0.648 and CopyKAT respectively obtained 0.309, as reported in Figure 1.9a.

As a further comparison, CopyKAT was executed using the non-malignant cells identified by SCEVAN. With this approach, CopyKAT obtained a much higher correlation with the ground truth. On the 26 samples of the GBM multiregional dataset, [112] it achieved a mean correlation of 0.33, as shown in Figure 1.9b. However, using the same classification of non-malignant cells, SCEVAN achieves a significantly higher correlation (p-value $1.3e^{-5}$) than CopyKAT The robustness of segmentation against misclassification of normal cells was also evaluated. Several cells were randomly removed from the reference control cells at 5% steps. Eight samples from the GBM multiregional dataset [112] were used. As shown in Figure 1.10, SCEVAN is robust to a high percentage of misclassified cells. The correlation of the copy number variation profile of the malignant cells with the ground truth remains stable for errors less than 60% and, in some cases, up to 95%. These results further confirm the robustness of the segmentation method for the misclassification of normal cells.

These data indicate that SCEVAN accurately infers DNA copy number profiles from high-throughput scRNA-seq data.

1.5.5 Computational Efficiency Comparison

SCEVAN is also particularly efficient since the main segmentation step is based on a greedy region-growing algorithm. To validate its performance in terms of computational efficiency, the classification step of the malignant cells and the segmentation step were compared separately. In the former case, the direct comparison of the execution times showed that SCEVAN

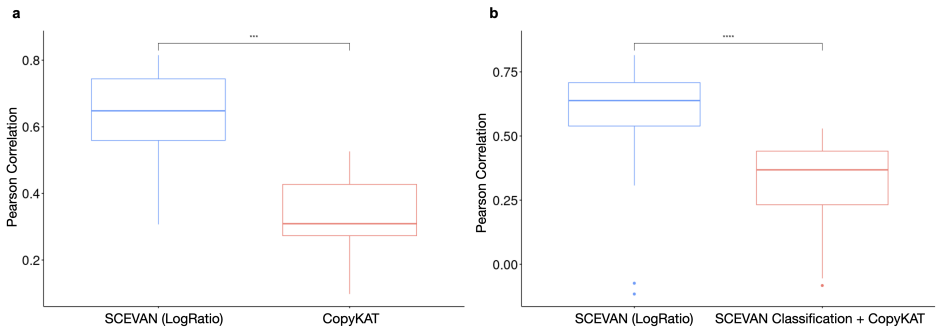


Figure 1.9. Comparison of the inferred copy number profile. In both **a** and **b** Box plots show the median as center, the lower and upper hinges that correspond to the 25th and the 75th percentile, and whiskers that extend to the smallest and largest value no more than 1,5*IQR. Values that stray more than 1.5*IQR upwards or downwards from the whiskers are considered potential outliers and represented with dots. Significance was computed by a two-sided Wilcoxon signed rank test (***: p-value <= 0.001, ****: p-value <= 0.0001). **a** A correlation comparison between the inferred copy number of SCEVAN and CopyKAT with ground truth from low-pass WGS data. The evaluation is only performed on samples in which CopyKAT has a good classification of tumor cells with an F1-score higher than 0.50. The 13 samples analyzed are S13P4, S3P4, S11P5, S13P6, S5P4, S13P5, S12P5, S6P2, S5P1, S6P5, S9P3, S1P2 and S3P8 from GBM multiregional dataset [112]. SCEVAN obtains a significantly higher correlation than CopyKAT, p-value $2.4e^{-4}$. **b** A comparison between the inferred copy number profile of SCEVAN and CopyKAT using as control cells the normal cells obtained from the SCEVAN classification. The evaluation is performed on 26 samples of GBM multiregional dataset [112] with the ground truth from low pass WGS data. SCEVAN obtains a significantly higher correlation than CopyKAT (p-value $1.3e^{-5}$), with a mean correlation of 0.57 against 0.33 of CopyKAT using SCEVAN classification as a reference.

is 2x to 7x faster (Figure 1.11a) in the discrimination phase between malignant and non-malignant cells. Afterward, comparing the time required for segmentation, on the multiregional GBM dataset, [112] SCEVAN is 2x faster than CopyKAT and 5x than inferCNV, instead for the Multiple Myeloma data [54], sequenced with 10x Genomics technology, CopyKAT becomes particularly slow, due to large number of cells. Specifically, as shown in Figure 1.11b, SCEVAN is 11x faster than inferCNV and 19x

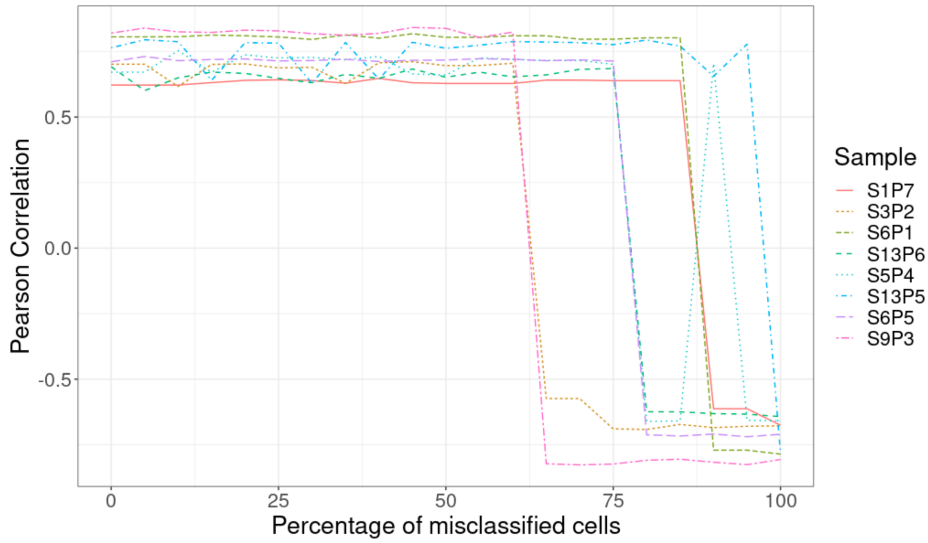


Figure 1.10. Correlation with ground truth varying misclassification errors. Correlation between the inferred copy number of SCEVAN with ground truth from low pass WGS data, to varying misclassification errors. The 8 samples are S1P7, SP32, S6P1, S13P6, S5P4, S13P5, S6P5 and S9P3 from GBM multiregional dataset [112].

than CopyKAT. These results show that the greedy segmentation algorithm implemented in SCEVAN is particularly efficient with respect to other methods for copy number inference from scRNA-seq.

1.6 Clonal substructure deconvolution

1.6.1 Intratumoral heterogeneity in Glioblastoma

Glioblastoma (GBM) is the most aggressive form of brain tumor. It is characterized by high heterogeneity, with several clonal and subclonal tumor cell populations, glioma stem cells, and an immuno-repressive tumor microenvironment [66, 16, 56]. SCEVAN can automatically infer clonal substructure from single-cell data by analyzing the clusters of the CNA matrix that show significantly different genomic alterations. As an appli-

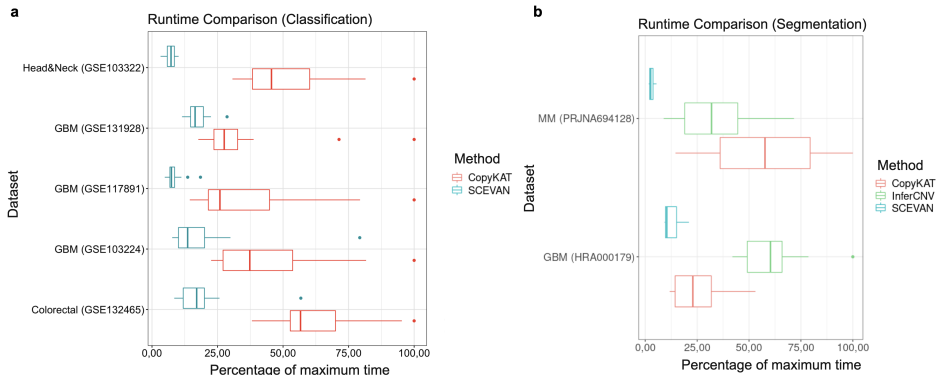


Figure 1.11. Runtime Benchmarking. In both **a** and **b**, Box plots show the median as the center, the lower and upper hinges that correspond to the 25th and the 75th percentile, and whiskers that extend to the smallest and largest value no more than 1,5*IQR. Values that stray more than 1.5*IQR upwards or downwards from the whiskers are considered potential outliers and represented with dots. **a** Classification time for each sample expressed as a percentage relative to the maximum time for each dataset Head and Neck Squamous Cell Carcinomas (GSE103322 [75]) n = 5,717 cells examined over 20 scRNA-seq independent experiments, GBM (GSE131928 [66]) n = 7,930 cells examined over 28 scRNA-seq independent experiments, GBM (GSE117891 [112]) n = 2,957 cells examined over 25 scRNA-seq independent experiments, GBM(GSE103224 [114]) n = 29,433 cells examined over 10 scRNA-seq independent experiments and Colorectal cancer (GSE132465 [49]) n = 47,285 cells examined over 23 scRNA-seq independent experiments **b** Runtime of copy number inference and segmentation for each dataset, expressed as a percentage relative to the maximum time, Multiple Myeloma (PRJNA694128 [54]) n = 34,204 cells examined over 7 scRNA-seq independent experiments and Glioblastoma (HRA000179 [112]) n = 2,957 cells examined over 25 scRNA-seq independent experiments.

cation of this approach, one of the samples reported in a recent study [66] is used, the MGH105 sample. SCEVAN identifies four sub-populations that have different alterations on chromosome 6 (Figure 1.13). Interestingly, whereas canonical scRNA-seq processing analyses could not reach the resolution for the identification of four subclones [66], instead the existence of these subclones had been previously described through the application

of DNA single-cell DNA methylation platforms [16].

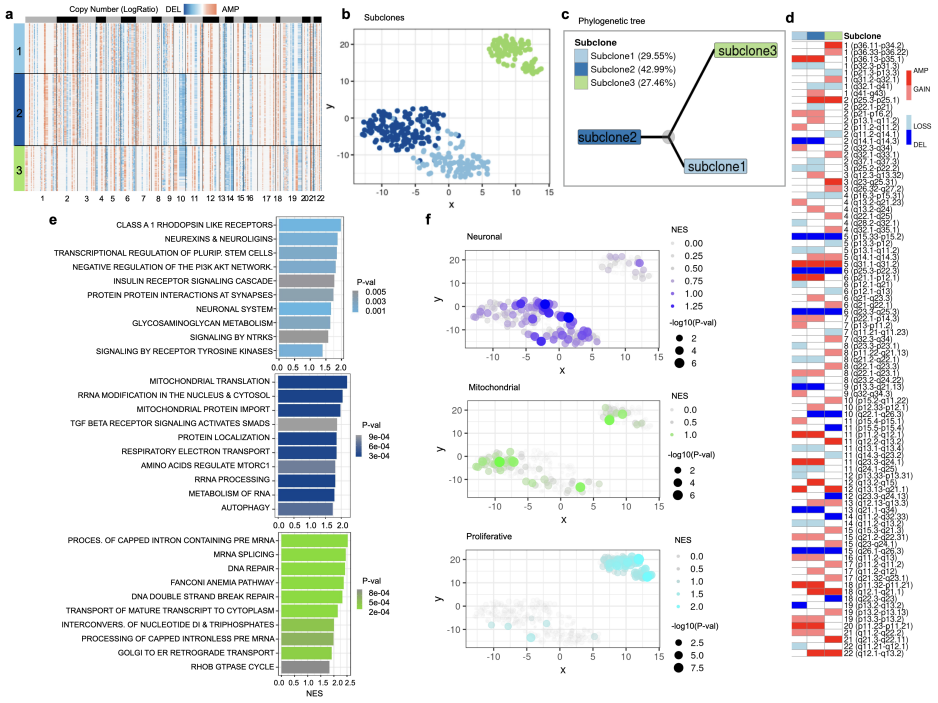


Figure 1.12. Deconvolution of the clonal substructure. **a** Clonal structure of sample BT1160 inferred by SCEVAN. **b** t-SNE plot of CNA matrix. **c** Inferred phylogenetic tree. **d** OncoPrint-like plot of BT1160 highlighting clone-specific alterations, shared alterations between and clonal alterations. **e** GSEA was performed on REACTOME[37] pathways for each subclone with a minimum size of 15 genes and a maximum size of 500 genes and with 10000 as number of permutations using the fgseaMultilevel function in the R package fgsea (v. 1.16), which calculates p-values based on an adaptive multilevel splitting Monte Carlo scheme. **f** NES and $-log_{10}(p\text{-value})$ per cell of GBM cellular states [28] computed by the Mann-Whitney-Wilcoxon single sample gene set test gene set implemented in the yaGST package [26]. Source data are provided as a Source Data file.

In sample BT1160, SCEVAN uncovers the presence of three tumor cell sub-populations, as shown in Figures 1.12a-b. Phylogenetic reconstruction of the clone tree shows two close clones (subclones 1 and 2)

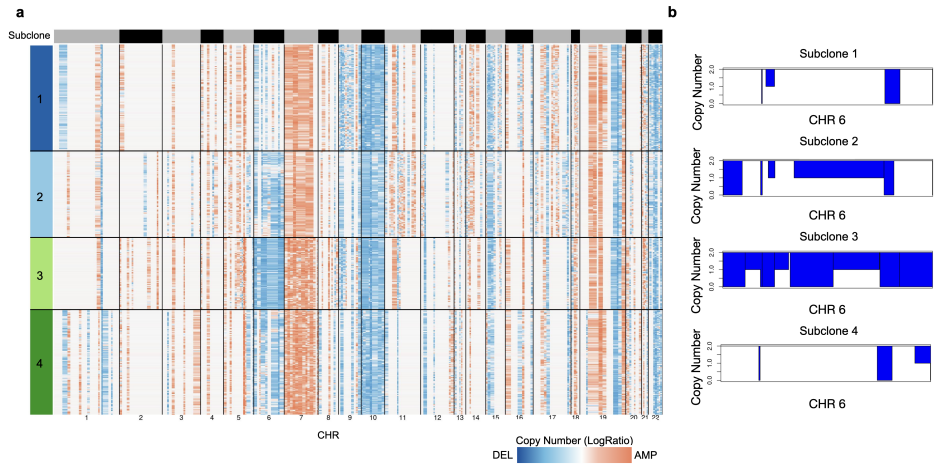


Figure 1.13. Subclonal copy number alterations of the chromosome. **a** Clonal structure of MGH105 [66] inferred by clustering single-cell copy number profiles by SCEVAN. **b** Inferred copy number in chromosome 6 for each subclone.

and a significantly far third subclone (Figure 1.12c). To better understand how individual clones fuel tumor growth and clonal selection, the reported alterations are investigated. SCEVAN identifies several truncal alterations, such as the amplification on Chr 5 (q23.2-q31.3), shared alterations, such as the deletion on Chr 10 (q22.1-q26.3), and subclone-specific alterations, such as the amplification in the green subpopulation on Chr 1 (q31.2-q32.1) and Chr 19 (q13.32-q13.33) (Figure 1.12d). Interestingly, subclone-specific functional analysis reveals a differential activation of pathways that resemble a recent metabolic classification of Glioblastoma [28]. Subclone 1 (lightblue) enriches pathways characteristic of the Neuronal subtype, subclone 2 (blue) has cells belonging to the Mitochondrial, and subclone 3 (green) contains cells with Proliferative/Progenitor subtype (Figure 1.12e). Indeed, this finding is confirmed by the enrichment of individual cells for every subtype (Figure 1.12f). The Proliferative/Progenitor subclone has several specific amplifications (1q21.3-q22, 1q31.2-1q32.1, 3q26.32-3q27.2, 4q32.1-4q35.1, 6p22.1, 8p11.22-8q21, 19q13.32-19q13.22). To identify drivers of the different cellular states, a differential analysis

between genes with genomic coordinates in regions of the subclone-specific alterations has been performed. The top differentially expressed gene lying in the alterations specific to the subclone 3 was the Ubiquitin-conjugating enzyme E2T (*UBE2T*) gene, which is significantly up-regulated (p-value $2.69e^{-43}$ log fold change 1.10) (Figure 1.14) enriching the activity of the pathway of DNA Repair. This gene encodes for the exclusive ubiquitin-conjugating enzyme (E2) that partners with the Fanconi Anemia (FA) ubiquitin ligase (E3). The E2T-FA complex is required for DNA inter-strand crosslink repair as the monoubiquitination event implemented by E2T is essential for the recruitment of downstream DNA repair factors by FA [33].

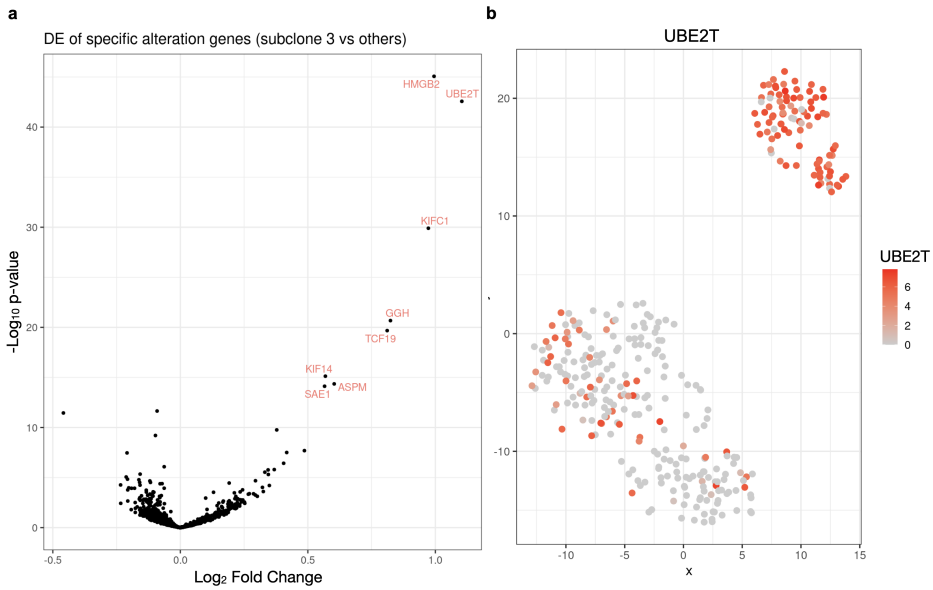


Figure 1.14. Top amplified and differential expressed gene of subclone. a Differential gene expression analysis of genes belonging to the specific amplifications of subclone 3, comparing subclone 3 against the others. Significance was computed by a two-sided t-test. **b** *UBE2T* expression on t-SNE plot of CNA matrix.

Furthermore, the analysis of copy number sub-structure can characterize the clonal status of specific tumor-associated genes. SCEVAN re-

veals that in samples BT1160 and MGH102, alterations of tumor suppressor genes *CDKN2A* and *PTEN* are subclonal (Figure 1.15). Indeed, in sample BT1160, the deletion on Chr 10 (q22.1-q26.3), containing *PTEN* (10q23.31), is shared between two out of three subclones, while in the remaining sub-population, this alteration is not present. Also, in the sample MGH102, the region 9p21.3 containing the gene *CDKN2A* is deleted in two of the four subclones. These results suggest that SCEVAN can resolve clonal copy number substructure in tumors from scRNA-seq data and identify subclonal differences and glioma-specific cancer states.

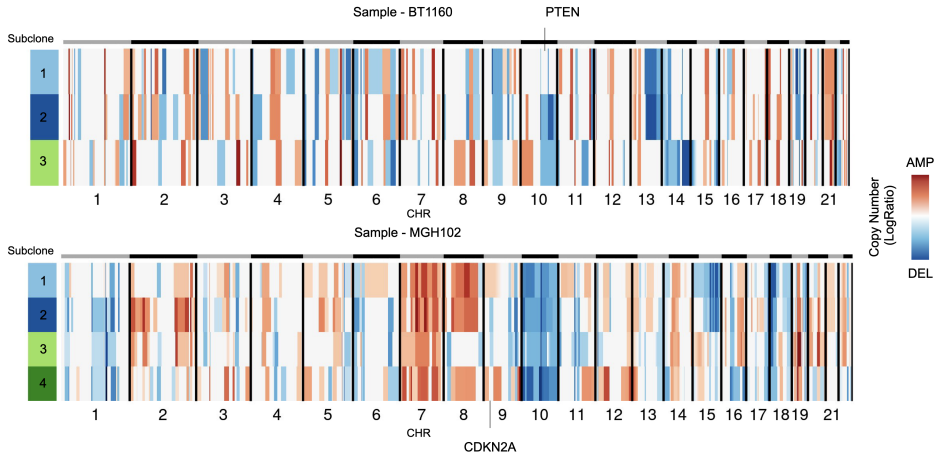


Figure 1.15. Tumour suppressor genes in the clonal substructure Compact representation of clonal structure inferred with SCEVAN of scRNA-seq samples BT1160 and MGH102 [66], in which the alterations containing tumor suppressor genes *PTEN* and *CDKN2A* are subclonal. Source data are provided as a Source Data file.

1.6.2 Clonal evolution in multiregional GBM tumor

Glioblastoma heterogeneity has also been investigated in the spatial and temporal axes [112, 9] because a single biopsy may not be informative of the whole tumor. Multiple biopsies allow to characterize the clonal architecture and evolutionary dynamics of GBM [50].

SCEVAN has been used for the evolutionary analysis of clonal structure for multiregional scRNA-seq samples of GBM [112]. For example, the patient GS1 with seven biopsies was analyzed, specifically with two taken at the tumor periphery and the remaining at the core of the tumor. The clonal analysis of each sample with SCEVAN allows to infer an evolutionary tree of the clones (Figure 1.16). Copy number alterations develop along several branches, and the peritumoral samples (P2/P3) are in a branch separated from the core samples, in which there is no amplification in chromosomes 4 and 8. Moreover, the amplification present on Chr 2 is clonal in peripheral samples and subclonal in some core samples (P1/P4/P7).

1.6.3 Clonal structure of primary and metastatic lymph

SCEVAN (and similar approaches) can address important questions, such as identifying similarities and differences between primary tumors and metastases. For this purpose, primary HNSCC tumors and corresponding lymph node metastases [75] were considered. Of the four considered cases, just one specific sample, the patient (HNSCC5), presented a different clonal structure between primary tumor and lymph node metastasis, particularly the absence of amplification of chromosome 7 (p22.3-p13) in the lymph node metastasis, as shown in Figure 1.17. Interestingly, this is the locus of Glycoprotein non-metastatic b (*GPNMB*), which is downregulated in lymph node metastasis (Figure 1.18). Furthermore, *GPNMB* increases tumor growth and metastasis in multiple contexts [53]. For the remaining patients (HNSCC20, HNSCC25, HNSCC26, HNSCC28), the clonal structure of the lymph node metastasis appeared to be the same as in the primary tumor. Therefore, a high correlation (Pearson correlation between 0.79 and 0.89) comparing the clonal profiles of the primary tumor and lymph node metastasis pairs is obtained. These data show that SCEVAN can be used to study the clonal evolution of metastatic cancer.

1.7 Findings

SCEVAN thus represents a powerful method that uses a joint segmentation algorithm to identify genomic copy number profiles from scRNA-seq

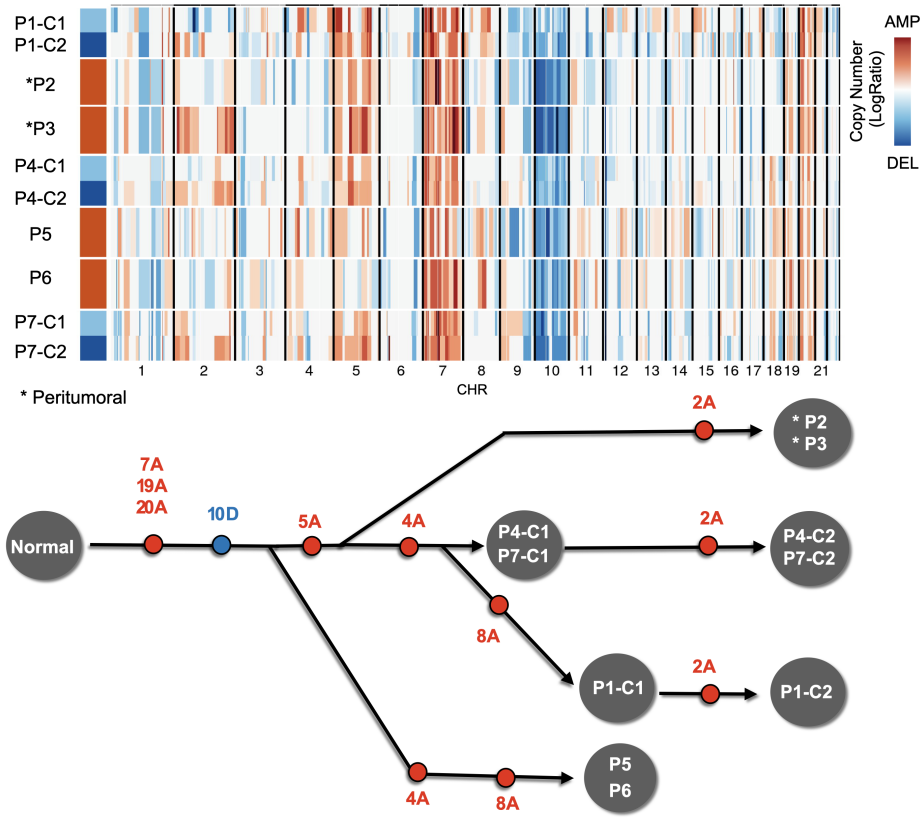


Figure 1.16. Temporal deconvolution of the clonal substructure. Compact representation of clonal structure inferred with SCEVAN of multiregional scRNA-seq samples of patient GS1 [112] and a phylogenetic tree deduced from clonal structure of the samples. Source data are provided as a Source Data file.

data. SCEVAN is more accurate and faster than state-of-the-art methods, making it ideal for delineating the clonal substructure in solid tumors and studying the temporal and geographic evolution of tumors. The evaluation of SCEVAN using annotated datasets of different tumor types has demonstrated its superior performance. It has been used in a large study integrat-

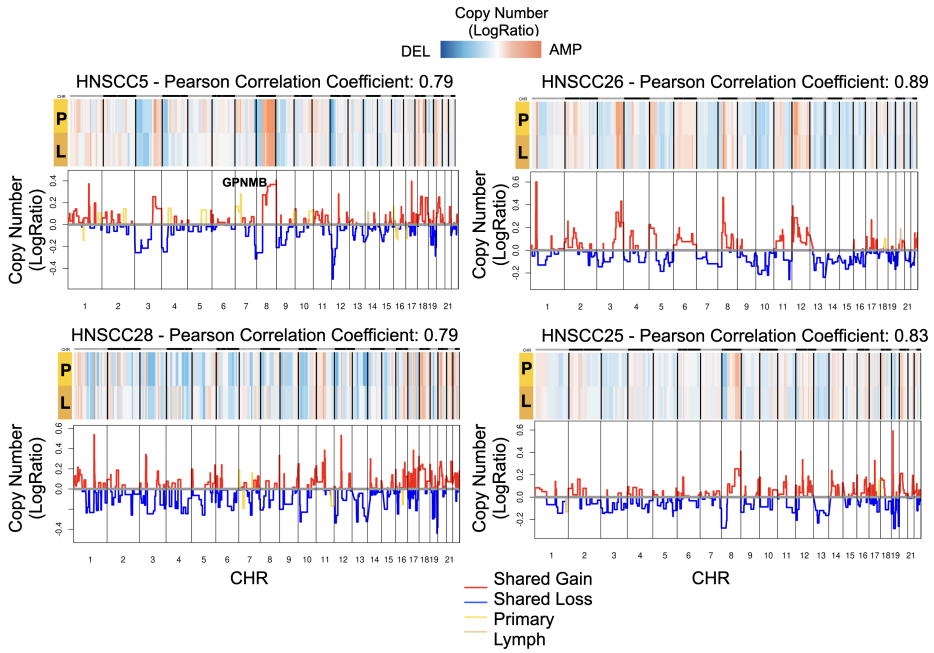


Figure 1.17. Clonal copy number comparison of matched Primary and Metastatic tumor. Copy number profile of Primary (P) and Metastatic Lymph nodes (L) from samples of Head and Neck cancer dataset (HNSCC5, HNSCC25, HNSCC26, HNSCC28) [75]. Source data are provided as a Source Data file.

ing millions of single cells from several datasets using different sequencing platforms (10X Chromium, Smart-seq2, Microwell-seq and STRT-C1).

SCEVAN can automatically discriminate between transformed cells and microenvironment cells, making it a valuable tool for identifying potential therapeutic targets. SCEVAN allowed the identification of UBE2T as a top amplified and differentially expressed gene in the PPR clone of glioma tumors. SCEVAN can also be used to study regional and temporal tumor evolution and to characterize the difference between primary and metastatic tumors.

However, it is important to note that SCEVAN relies on the assumption that aneuploidy can identify cancer cells. While it is a highly effective

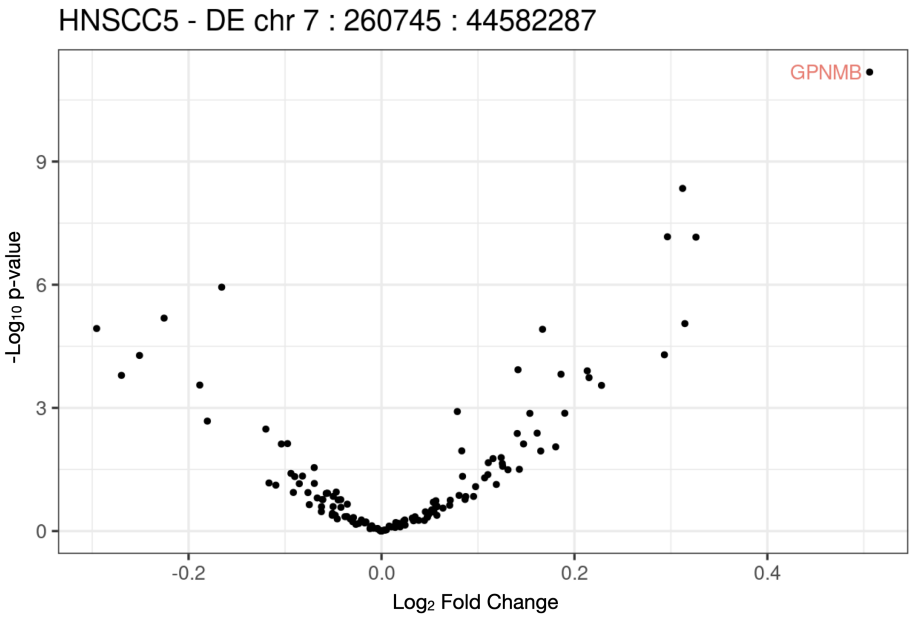


Figure 1.18. Differential analysis in specific copy number alteration. Differential analysis of genes belonging to the specific amplification on chromosome 7, between the gene expression of the primary tumor against the lymph node metastasis. Significance was computed by a two-sided t-test.

method for many types of cancers, it may not be well suited for certain types of cancers, such as liquid cancers, pediatric cancers, and Ependymomas, which may have a minimal number of genomic alterations. Thus, SCEVAN approach (and similar) may not be suited in this case.

Chapter 2

Liquid Biopsy

Detecting, characterizing, and monitoring cancer with liquid biopsy can transform cancer care from early detection to resistance and response prediction and a better definition of personalized therapies. Liquid biopsy is based on the analysis of cell free DNA (cfDNA), opening the possibility of taking repeated blood samples, consequently allowing the changes in cfDNA to be traced during the natural course of the disease or during cancer treatment for obtaining diagnostic information that has previously only been obtainable through invasive biopsies. On the contrary, acquiring tumor biopsies by invasive methods poses a significant challenge in accurate tumor profiling and monitoring. Extracellular DNA circulates in the blood due to cell death, which can result from normal cellular processes or specific diseases [81, 98]. For healthy individuals, most of the cfDNA in the plasma originates from blood cells [90]. Organs or tissues affected by a particular disease lead to increased cell death, which results in a more abundant presence of cell free DNA from that specific organ or tissue in the blood. Standard technologies based on next-generation sequencing adopt a "tumor-informed" approach by analyzing parts of the genome at very high depth to discover the presence of target somatic alterations. Without prior information or hotspot mutations, different "blind" genome-wide approaches are needed. Since multiple tissues contribute to the circulating DNA, identifying the tissue of origin is the primary diagnostic strategy to uncover which organ generated disease-associated DNA molecules and, eventually, large amounts of DNA [81]. The tissue of origin must be based

on features other than the DNA sequence itself, as all cells contain the same genome. Within this context, the epigenetic footprints of cfDNA arise as a critical factor for successfully exploiting all potential of liquid biopsies [55, 88](Figure 2.1). The epigenetic status is highly tissue-specific and can accurately distinguish cancer types [32] and even specific cancer subtypes with clinical relevance [13]. Multiple epigenetic patterns inform tissue of origin in cfDNA, including the nucleosome organization [88], end-motifs distribution [71], fragmentation pattern [20], and DNA methylation signatures [86].

2.1 Epigenetics features

The cancer epigenome is characterized by global changes in DNA methylation and histone modification patterns [84].DNA methylation refers to the addition of methyl groups to the DNA molecule. DNA methylation is an important epigenetic regulation because it causes changes in gene expression without altering the DNA sequence; in fact, when methylation occurs in the promoter of a gene, it typically acts to repress gene transcription [105]. Another epigenetic mechanism is related to the nucleosome organization, which is the basic unit of chromatin, consisting of DNA wrapped on histone proteins. The organization of nucleosomes refers to how DNA is packaged within these units. Different cell types have unique patterns of nucleosome organization, which influence the accessibility of DNA and, consequently, gene regulation and expression [86]. Epigenetic signatures related to nucleosome organization, such as fragment size and position coverage and DNA methylation, as detailed below, bring with them various evidence of cancer-specific characteristics that can be used for cancer detection and monitoring.

2.1.1 Fragment lengths

Fragmentation pattern refers to the size distribution of cfDNA fragments. The peaks of fragment length distribution, as shown in Figure 2.3, reflect precisely the organization of nucleosomes. There is a significant difference in fragment size distribution between healthy and cancer samples [20]. In the organization of nucleosomes, as known (Figure 2.2A), the

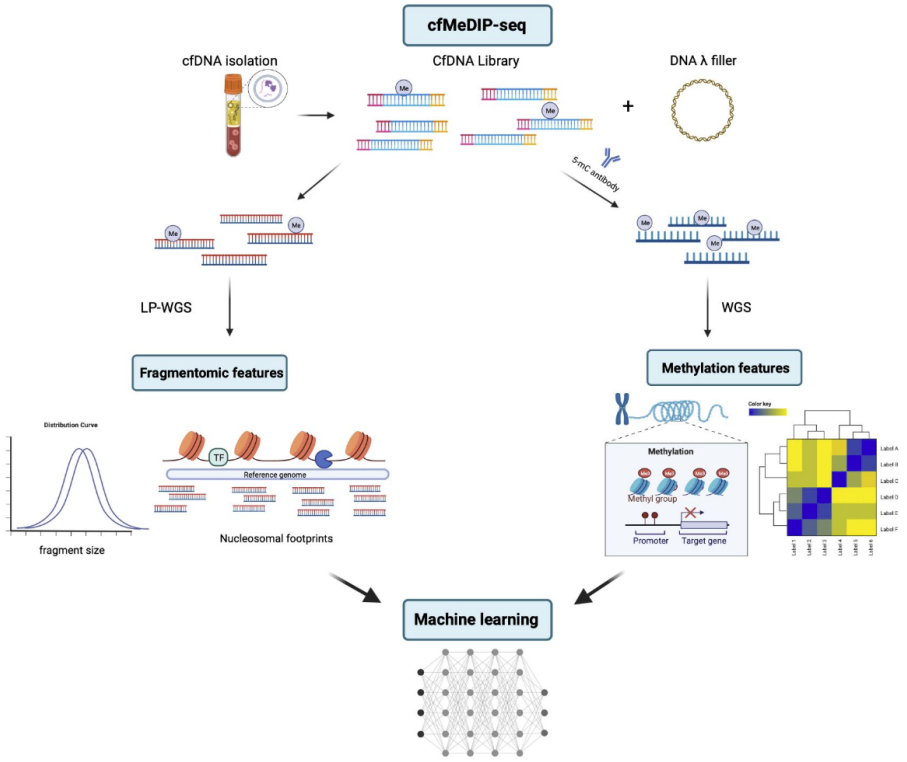


Figure 2.1. A clinical grade cfDNA protocol for cancer detection and monitoring using the Fate-AI technology. The cfMeDIP-Seq protocol is a high-sensitivity method capable of assessing methylation status from low amounts of input cfDNA through the addition of a DNA filler, a pool of methylated and unmethylated PCR amplicons, used as a carrier for the immunoprecipitation reaction with anti-methylcytosine antibodies. The cell free DNA library is isolated from 1 ml of plasma, and the sample is divided for the generation of two different libraries. The library subjected to the immunoprecipitation reaction with 5-mC antibodies is used for methylation analysis. The other library is sequenced by low-pass whole genome sequencing and used for fragmentomics analysis. The combination of fragmentomics and methylation features was used for the development of a classifier, which is used by a machine learning model for cancer detection and tissue of origin classification.

nucleosome core is comprised of approximately 145bp of DNA wrapped around an octamer of histone proteins, constructed from two copies of each histone. Each nucleosome core is connected to an adjacent nucleosome core through a segment of linker DNA, approximately 20 bp of this linker DNA is typically found in association with the linker histone H1, increasing the nuclease protection of the core particle to 167 bp. The nucleosome core, together with the linker histone, is called the chromatosome. Adding the remaining linker DNA to the chromatosome completes the nucleosome, forming the chromatin polymer with a repeat length ranging from 160 to 240 bp with a peak of 197bp.

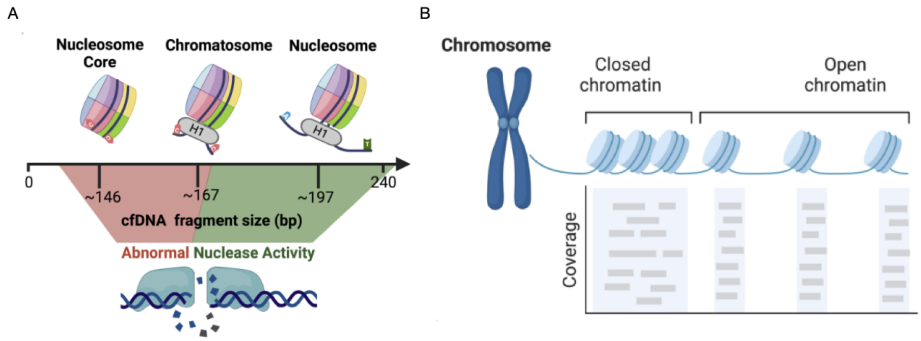


Figure 2.2. Nucleosome organization. A. The nucleosome core comprises about 145 base pairs of DNA wrapped around a histone octamer. Adjacent nucleosome cores are connected via a segment of linker DNA, increasing nuclease protection to approximately 167 base pairs (chromatosome). The addition of the remaining linker DNA completes the nucleosome, resulting in the chromatin polymer, with a repeat length ranging from 160 to 240 base. B. An organized nucleosome structure results in decreased sequencing coverage, suggesting DNA degradation at the exposed binding site. Conversely, high coverage peaks are evident at adjacent protected positions.

As shown in Figure 2.3, there is a significant difference in fragment size distribution between healthy and cancer samples around the peaks described above, in particular, an increase of fragments with a length related to the nucleosome core occurs in the patients with cancer disease, and inversely an increase of fragments from chromatosome to the complete nucleosome in the patients without disease. This different distribution, as

demonstrated in several previous works ([29] [71] [83]), is due to abnormal nuclease activity, with a particular interest in the endonuclease enzyme DNASE1L3, which in several experiments on mice or tumor samples causes this different fragment distribution.

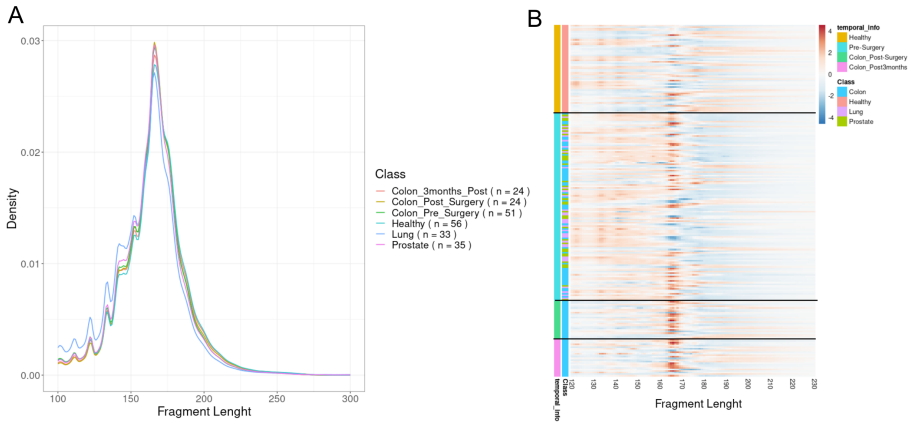


Figure 2.3. Fragment Distribution. A. Fragment lenght distribution in colon, lung, prostate and healthy samples. B. Heatmap of the z-score of the density of each fragment length calculated against a reference of the density of each fragment length of all healthy samples.

2.1.2 End-motif

End-motifs refer to the specific DNA sequences or motifs at the ends of cfDNA fragments, for them too, the abnormal nuclease activity affects a different distribution. End-motifs show a different distribution among tissues, also in this case related to the abnormal activity of endonuclease enzyme DNASE1L3 [71]. In Figure 2.4, the six end-motifs (CCCA, CCAG, CCTG, TAAA, AAAA, and TTTT) identified as differential between liquid biopsies from healthy patients and patients with cancer are analyzed [71]. As shown, they are significantly differential, compared to healthy samples, depending on the patient's tumor type. The different distribution compared to healthy samples and between the different tumor classes themselves indicate that the signature derived on the basis of end-motifs

may allow the identification of cancer patients and discrimination of tissue of origin from the liquid biopsy.

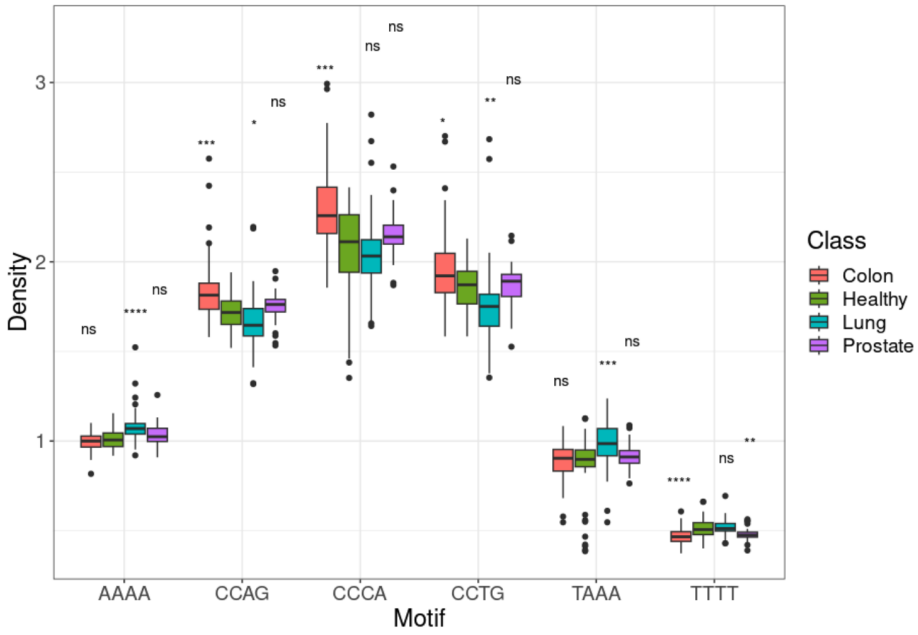


Figure 2.4. End-motif. Distribution of six end-motifs (CCCA, CCAG, CCTG, TAAA, AAAA, and TTTT)[71] in colon, lung, prostate and healthy samples.

2.1.3 Nucleosomal footprinting

At genomic sites that are accessible, such as actively bound transcription factor binding sites (TFBSs) and open chromatin regions, nucleosomes are arranged in an organized fashion to facilitate access to DNA-binding proteins. This organized nucleosome structure leads to reduced sequencing coverage, indicating DNA degradation at the exposed binding site, with elevated coverage peaks observed at the adjacent protected locations (Figure 2.2B). For this reason, the analysis of coverage profiles around TFBSs established a distinctive quantitative feature in the ±30 bp window (central

coverage), which distinguishes the accessibility of a site [23]. This feature serves as a measure to quantify transcription factor activity. Transcription factors (TFs) are crucial in the dynamic regulation of gene expression. Recent advances in the understanding of cellular gene expression programs highlight the significant impact of gene expression dysregulation in diseases, including cancer. Mutations affecting TFs, contribute to numerous diseases, indeed the dysregulation of the activation state of specific TFs is linked to various types of cancer, with many oncogenes and tumor suppressor genes (es. TP53) acting as TFs [103].

Figure 2.5 shows the inferred TFs activity based on the central coverage of TFBS [23] in colon and lung cancer. Among the TFs with significant differential activity between healthy and cancer samples, there are ATF3, BATF3, FOS, FOSL2, JUN, and JUND which are subunits of the transcription factor AP-1 (activator protein1). The oncogenic role of AP-1 is recognized in several cancer types. AP-1 plays a key role in the mechanism of increased proliferation and resistance to cell death, progression, and invasion [107]. In particular, AP-1 is one of the most important oncogenic TFs in colorectal cancer [110] and also plays a key role in lung cancer progression [47].

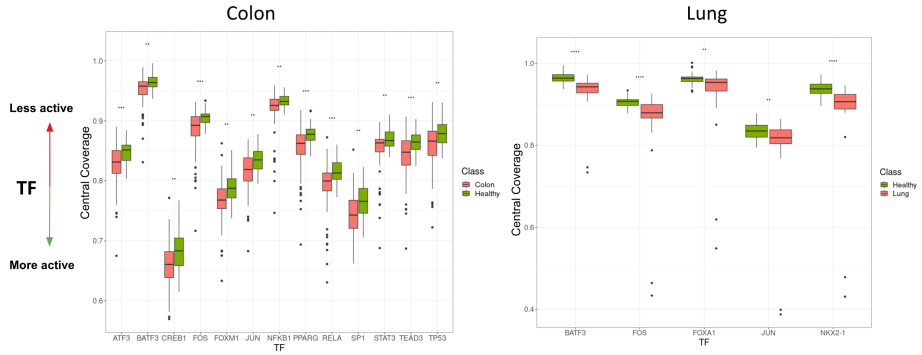


Figure 2.5. Trascription Factor Activity. Transcription Factor Activity inferred from central coverage of TFBSs. In the left panel the differential TFs in colon cancer respect to healthy samples. In the right panel the differential TFs in lung cancer respect to healthy samples.

In addition, based on the coverage, it is possible to analyze the DNase I hypersensitive sites (DHS), regions of the genome where chromatin has lost its condensed structure, exposing DNA and making it accessible due to the sensitivity to cleavage by the DNase I enzyme. In particular, since the Hematopoietic cells are the stem cells that give origin to other blood cells, they represent the main source of cfDNA in healthy controls. Analyzing coverage in the specific DHS sites of hematopoietic cells [8], as expected, there is a loss of coverage and so higher accessibility for Healthy and even post-surgery Colon samples (Figure 2.6).

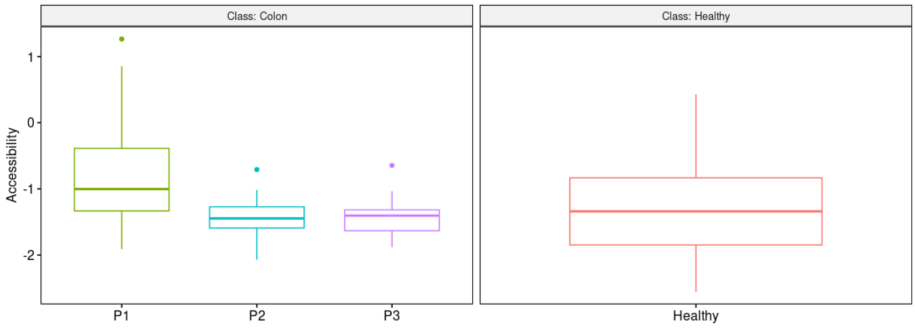


Figure 2.6. Hematopoietic DHS. Coverage analysis in the Hematopoietic specific DNase I hypersensitive sites (DHS). Comparison of temporal Colon cancer liquid biopsy (blood samples that were collected pre-surgery (P1), after one month (P2), and three months from the surgery (P3)) and from Healthy patients.

2.1.4 Methylation

In cancer, DNA methylation leads to silencing of tumor suppressor genes or activation of oncogenes [48]. Aberrant DNA methylation is a hallmark of cancer and is associated with tumor initiation, progression, and metastasis, DNA methylation patterns so can be used as a diagnostic and prognostic marker for cancer and is also a promising target for cancer therapy [48]. The recently developed technology, cell free methylated DNA immunoprecipitation sequencing (cfMeDIP-seq) [85], is an adapted version of MeDIP-seq to allow immunoprecipitation reaction of a low amount of

input cfDNA. This technology allows its use in liquid biopsy sequencing data, enabling extraction of methylation patterns.

The cfMeDIP-seq data extracted from plasma were typically analyzed with the following pipeline[67]: the methylation profile in each 300bp region of the genome is extracted, and the profile is compared between patients with cancer and between patients in the control group (healthy donors). Regions with FDR < 0.01 and the absolute value of log2 fold change > 1 were selected as differentially methylated regions (DMRs). The top 300 DMRs were used to train a classifier capable of discriminating the methylation profiles of patients with and without cancer.

2.2 Fate-AI

2.2.1 Data pre-processing

Blood samples from treatment-naive patients with Prostate cancer (35 patients) and Colon cancer (51 patients) were collected at Istituto Oncologico del Mediterraneo IOM (Catania, Italy), for 24 of the latter patients two further blood samples were collected after one month and three months from the surgery. Blood samples from 33 treatment-naive patients with lung cancer were collected at Azienda Ospedaliera Universitaria "Luigi Vanvitelli" (Caserta, Italy) and finally, blood samples from 62 healthy donors were used as controls. Low-pass whole genome sequencing and cfMeDIP-seq were performed for each sample. The cfMeDIP-Seq protocol is a highly sensitive method that can determine methylation status even from low amounts of cfDNA input. This is done through the addition of a DNA filler, which is a pool of methylated and unmethylated PCR amplicons. It is used as a carrier for a successful immunoprecipitation reaction with 5-methylcytosine antibodies. The library is isolated from 1mL of plasma, after which the sample is divided for the generation of two different libraries. The library that undergoes the immunoprecipitation reaction with 5-mC antibodies is used for methylome analysis. On the other hand, the other library is sequenced by low-pass whole genome sequencing and used for the study of fragmentomics. All libraries are sequenced on Novaseq 6000 (Illumina). The sequencing reads obtained from the Low pass whole genome sequencing (LPWGS) of the liquid biopsy were aligned with re-

spect to human genome hg38 using Sentieon bwa-mem (0.7.17-r1188), then duplicate reads were removed and finally, the indel realignment and base recalibration were performed using Sentieon tools v. 202112. cfMeDIP-seq are processed using the pipeline described by Nuzzo et al. [67].

2.2.2 2D-Fragmentomics Profile

The fragmentomics signatures released from circulating tumor DNA within the liquid biopsy, allow to identify the presence of the disease and since they reflect the genomic alteration of the primary tumor [20], for this reason Fate-AI extracts along the genome decomposed into different regions these features based on the fragment length and end-motif that can discriminate between circulating tumor and circulating free DNA fragments. In order to determine the best resolution of the fragmentomics profile, it was computed at different resolutions by varying the bin size (0.5 Mb, 1Mb, 1.5 Mb, 2Mb, 2.5Mb, 3Mb, 3.5Mb, 4Mb, 4.5Mb, 5Mb), which is the size of the region of the genome in which the fragmentomics features are extracted. The correlation of each feature with a median copy number profile of Colon cancer was computed from the TCGA database while varying the size. A bin size of 3Mb was chosen as the optimal resolution for the fragmentomics profile, exhibiting the highest median correlation.

The genome is subdivided into non-overlapping regions of 3 million bases, and in each of these regions, 19 features (Figure 2.8) based on the DNA fragment size, coverage and end-motif are extracted.

The following features are calculated based on the length of the fragments in each region: Median absolute deviation, Standard deviation, Coefficient of variation, Shannon entropy, and Mean.

Coverage based on the different fragment sizes is then extracted in each region and normalized against the median of all regions: coverage (all fragments), coverage nucleosome core (>=140bp & <= 159bp), coverage chromatosome (>=160bp & <= 170bp), coverage nucleosome (>=171bp & <= 240bp).

The ratio of different coverage in each region is then calculated: ratio coverage nucleosome core (>=140bp & <= 159bp)/coverage nucleosome (>=171bp & <= 240bp), ratio coverage chromatosome (>=160bp & <= 170bp)/nucleosome, ratio nucleosome core + chromatosome (>=140bp & <= 170bp)/nucleosome (>=171bp & <= 240bp) and mononucleosome

Short-Long Ratio (Short <= 120bp — Long >=140bp & <= 250bp).

Finally, the percentage presence of six known 4-mer end-motifs typically altered in cancer (CCCA, CCAG, CCTG, TAAA, AAAA, and TTTT) [71] out of the total end-motifs is extracted in each region. The analysis involved extracting the initial four $5'$ nucleotides from the reference genomic sequence for every read. The first four $3'$ nucleotides were omitted due to potential alterations caused by end repair during library preparation, which might not accurately represent the native genomic sequence [71].

The fragmentomics profile along the genome extracted from the Low pass whole genome sequencing (LPWGS) data of the liquid biopsy reflects genomic alterations in the primary tumor. Indeed, as shown in Figure 2.8 the different features reflect the profile of the copy number obtained from the matched tumor tissue, the comparison was performed using the public dataset containing the liquid biopsy and the matched tissue biopsy of pediatric patients with Ewing sarcoma [8].

2.2.3 Auto-Encoder Model

A Convolutional Auto-Encoder was designed to learn efficient representations of the 2D-Fragmentomics profiles by encoding them in a low-dimensional latent space (Encoder) and then decoding them in their original form (Decoder). In the latent space by reducing dimensionality, the auto-encoder forces the network to learn a compressed but informative representation of the fragmentomics profile obtained from the liquid biopsy. Convolutional layers are used in both the encoder and decoder to extract and reconstruct features from the data.

This architecture, shown in Figure 2.9, is particularly suitable for processing grid-like data, such as the 2D-Fragmentomics profile, as they use learnable filters (kernels) to pass over the input data, capturing local patterns and features. The representations of the latent space are used to feed a shallow classifier for tumor detection and tissue of origin classification.

The model starts with an input layer, which takes the 2D-Fragmentomics profile $b \times f$ where b is the number of bins (938) and f is the number of features (19). The encoder section of this model is responsible for reducing the dimensionality of the input data and capturing essential features. It consists of three convolutional layers. The first convolutional layer has 16 filters. It uses a 1x3 kernel, meaning it scans the input data vertically with

a 1x3 stride, capturing patterns in the same region (bin) of the genome. The activation function used is ReLU, which introduces non-linearity. The second convolutional layer has eight filters with the same kernel size as the previous layer. It also employs ReLU activation and the same stride. The final encoder layer employs four filters. It uses a larger 2x2 kernel, capturing broader features in the data. Like the previous layers, it uses ReLU activation, 'same' padding, and strides of (2, 2). After the encoding layers, a flattening layer reshapes the output into a flat vector. This represents the feature vector in the embedding space of the Fragmentomics profile, which can be used as input for a subsequent classifier. Afterward, there is the decoder portion of the model aims to reconstruct the original data from the compressed representation generated by the encoder. In this section, is first present a Reshaping Layer in which the flattened output from the encoder is reshaped to match the dimensions of the last convolutional layer's output. Then, there are three Transpose Convolutional Layers that perform the reverse operation of the encoder, effectively "upsampling" the compressed representation. The parameters for these layers are similar to their corresponding encoder layers but in reverse order. They use ReLU activation functions and 'same' padding to restore the data's dimensions. Lastly, the final Convolutional Layer uses a 1x2 kernel, and its purpose is to produce the final output. It applies a sigmoid activation function to generate values between 0 and 1, making it suitable for binary data. The output of this layer represents the model's attempt to reconstruct the original input data. After decoding, a Cropping2D layer is applied to remove any extra padding introduced during the convolution and deconvolution operations. This step ensures that the model's output matches the original input size. The model is trained using the Adam optimizer with a learning rate of 0.001. The loss function used for training is the binary cross-entropy, which measures the dissimilarity between the input and the predicted output. Training is carried out over 100 epochs with a batch size of 4. The data is shuffled to introduce randomness during training. An early stopping callback is employed to monitor the validation loss and potentially stop training early if the model stops improving. In summary, the Fate-AI model is a convolutional autoencoder designed for the compression and reconstruction of the Fragmentomics profile. The encoder compresses input data through convolutional layers, the flattened output

is decoded using transposed convolutional layers, and the final output is cropped to match the input size. Finally, the latent space obtained from the trained convolutional auto-encoder was used as input for a Logistic Regression-based classifier that can discriminate the fragmentomics profiles of patients with and without cancer with high accuracy.

2.3 Fate-AI benchmark

Fate-AI was compared on the task of early cancer detection with the two methods DELFI [20] and GRIFFIN [23], representing the state-of-the-art in the liquid biopsy field based on fragmentomics. DELFI [20] based only on coverage relative to fragment size, uses as feature the GC-corrected total (>=151bp and <=220bp) and short fragment coverage (>=100bp and <=150bp) for all 504 bins of 5Mb were centered and scaled for each sample to have mean zero and unit standard deviation and a stochastic gradient boosting model (GBM). GRIFFIN [23], instead based on the nucleosome profiling of 270 TFs (30,000 TFBSs for each TF), uses as features the coverage in the window between ±30bp for which lower values represent increased accessibility (central coverage), the coverage in a window between ±1000bp (Mean coverage) and overall nucleosome peak amplitude calculated using Fast Fourier transform (Amplitude). Then PCA is calculated on the previous features, and logistic regression is used for prediction, giving as input the first components expressing 80% of the variance.

In addition, a benchmark is conducted for the early detection task of Fate-AI based on the combination of features extracted from the two sequencing (LPWGS and cfMeDIP-seq) of the same sample. The embedding of the 2D profile of fragmentomics computed on the Low pass whole genome sequencing (LPWGS) data and the 300 DMRs computed in each training set from cfMeDIP-seq data are concatenated. The concatenated vector was used as input for a Logistic Regression-based classifier.

Leave-one-out cross-validation (LOOCV) is adopted to evaluate the prediction performance and determine the Area Under Curve (AUC).

Furthermore, the use of Fate-AI is also verified for identifying the tissue of origin and performing the MRD task.

2.3.1 Early detection of Cancer

Fate-AI reaches superior performance in the early detection of cancer compared to healthy samples from the liquid biopsy. In colon cancer, it obtains an AUC of 0.973, sensitivity of 96%, and specificity of 89%, and AUC of 0.93 using only fragmentomics, versus an AUC of 0.852 and 0.82 for DELFI [20] and GRIFFIN [23], respectively (Figure 2.10A). For lung cancer detection, it obtains an AUC of 0.985, a sensitivity of 98%, and a specificity of 94% and AUC of 0.93 using only fragmentomics, versus an AUC of 0.94 and 0.88 for DELFI [20] and GRIFFIN [23], respectively (Figure 2.10B). High accuracy is also observed in prostate cancer, in particular using the integration of fragmentomics and methylation that obtains an AUC of 0.99 with a sensitivity of 94% and specificity of 98% (Figure 2.10C), as opposed to an AUC of 0.78 and 0.94 for DELFI [20] and GRIFFIN [23]. The Fate-AI score, in addition, is effective in disease staging in colon cancer with a correlation of 0.64 between score and stage (Figure 2.10D).

2.3.2 Tissue-of-origin Identification

One of the main applications of liquid biopsy is the identification of the tissue of origin, which involves determining the primary site of a tumor. This information is crucial for guiding therapeutic decisions and predicting patient outcomes. By analyzing specific genetic alterations or expression patterns in ctDNA, is possible to infer the tissue of origin and provide valuable insights into personalized cancer management.

Fate-AI can also be used to identify the tissue of origin accurately, as shown in Figure 2.11. The correct identification of the primary tumor tissue is verified with a binary classification task comparing one class against all other available tumor classes (Colon, Lung, Prostate) each time. Fate-AI obtained good identification accuracy, particularly for the lung cancer class with an AUC of 0.97.

2.3.3 Minimal Residual Disease

Minimal residual disease (MRD) refers to the small number of cancer cells that may remain in the body after treatment. Detection of MRD is crucial for assessing response to treatment and predicting the likelihood

of recurrence. Liquid biopsies are emerging as a less invasive and more accessible alternative to traditional methods for assessing MRD. Their non-invasiveness allows continuous monitoring at such a frequency as to have a clear temporal view of the progress of therapy and to recognize possible relapses in time.

For Colon cancer, some temporal samples are employed, specifically for 24 patients the blood samples that were collected pre-surgery (P1), after one month (P2), and three months from the surgery (P3). Training the model on healthy, pre-surgery samples allows Fate-AI to predict a probability score of the sample belonging to a tumor sample and consequently predict a response to treatment/surgery, etc. The score is then obtained by showing a later time sample (post-intervention samples) to the Fate-AI, which allows monitoring of the disease. The score, as shown in Figure 2.12, in almost all samples decreases after one month (P2 samples), but in some cases, an increase is observed in P3 samples, in some cases consistent with known clinical information of progression or metastasis or new primary tumor occurring after three months (Figure 2.13). Score increase also occurs for some patients without currently known information on progression.

2.4 Findings

The proposed Fate-AI method has considerable relevance, as it has the potential to revolutionize early cancer detection and therapy monitoring, which can have a direct impact on the survival rates and quality of life of cancer patients. The proposed solution introduces an innovative method that involves the analysis of cfDNA, which has emerged as a promising non-invasive strategy for cancer diagnostics. The epigenetic modifications of cfDNA can indicate specific cancer types and stages, which makes it a valuable tool for early detection of cancer and therapy monitoring.

Early detection of cancer has been shown to increase survival rates, reduce treatment costs, and positively affect the lives of both patients and their families. The advancements proposed in Fate-AI could lead to more effective, less invasive, and cost-effective cancer diagnostics, thus enhancing patient outcomes and reducing mortality rates. This could have a far-reaching impact on society as it would allow for more accurate and efficient diagnosis and treatment of cancer, which could help save countless

lives and improve the quality of life of cancer patients.

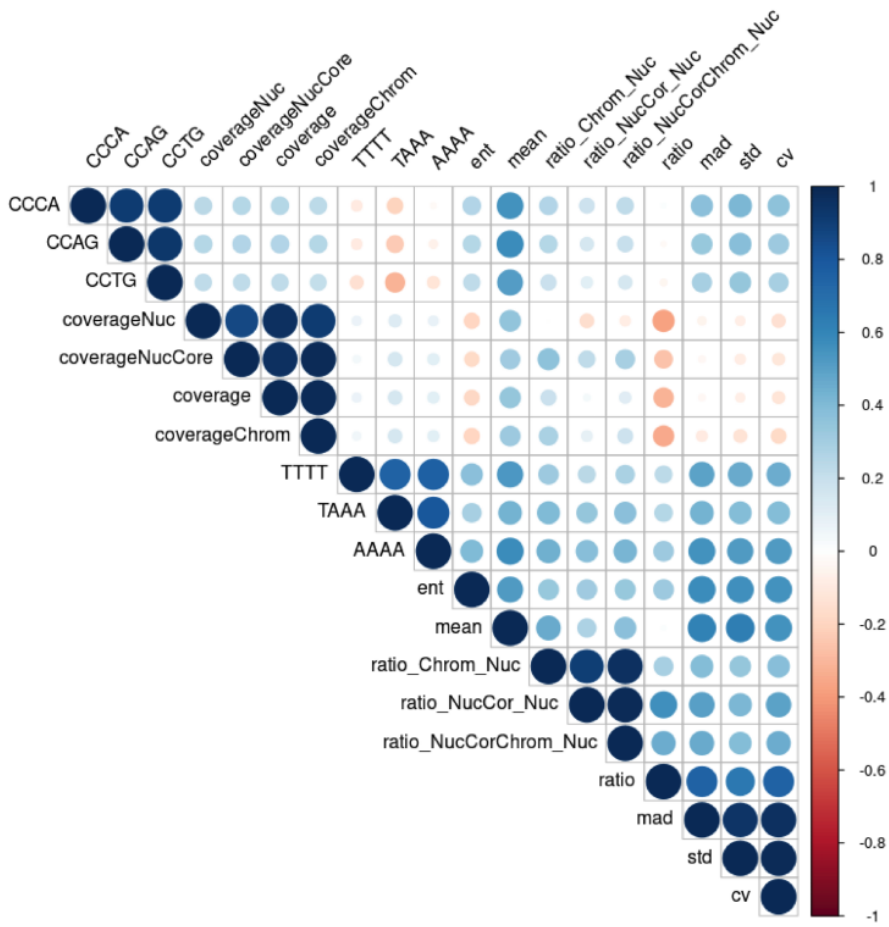


Figure 2.7. Correlation between Fragmentomics features used in Fate-AI. Fate-AI subdivides the genome into non-overlapping regions of 3 million bases. In each of these regions, a set of 19 features based on the DNA fragment size and end-motif are extracted: Median absolute deviation, Standard deviation, Coefficient of variation, Shannon entropy, Mean, coverage, coverage nucleosome core, coverage chromatosome, coverage nucleosome, ratio nucleosome-core/nucleosome, ratio chromatosome/nucleosome, ratio nucleosome core + chromatosome/nucleosome, mononucleosome Short-Long Ratio, density of specific end-motifs (CCCA, CCAG, CCTG, TAAA, AAAA, and TTTT).

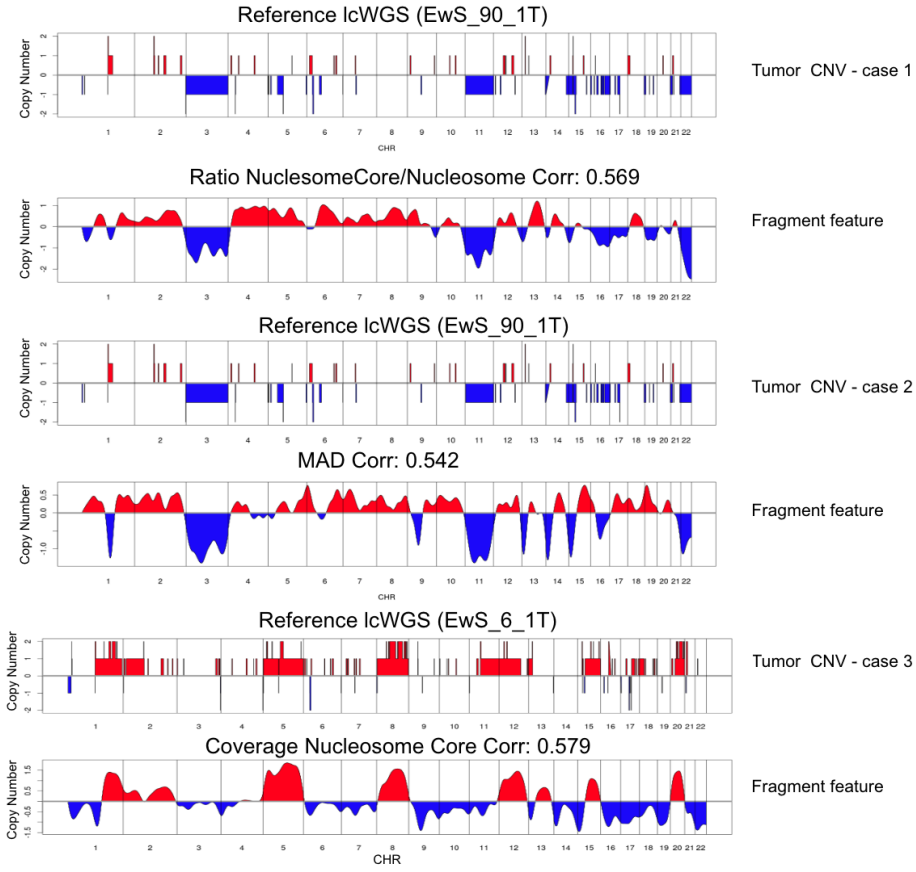


Figure 2.8. Correlation Fragmentomics features and copy number profile from tissue. Three copy number profiles of Ewing Sarcoma, fragmentomics features reflect changes in copy number profile.

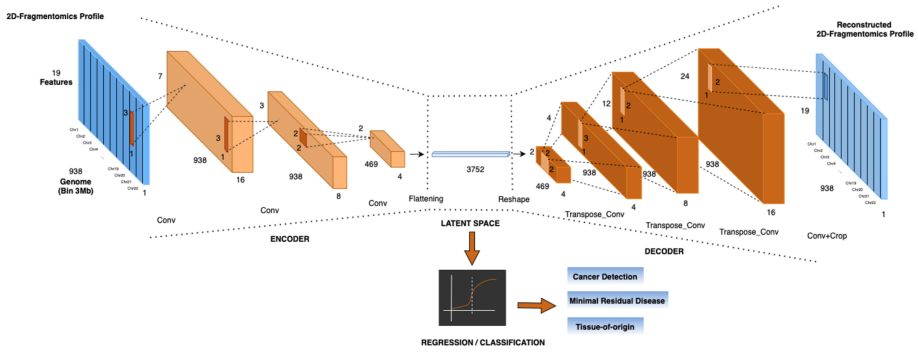


Figure 2.9. Fate-AI Architecture. The model takes in input the 2D-Fragmentomics profile along the genome. The encoder section consists of three convolutional layers. The first layer has 16 filters, a 1x3 kernel, and employs ReLU activation. It scans the input data vertically with a 1x3 stride, capturing patterns in the same region (bin) of the genome. The second convolutional layer has 8 filters with the same kernel size as the previous layer. It also employs ReLU activation and the same stride. The final encoder layer employs four filters and instead uses a larger 2x2 kernel, capturing broader features in the data. In the decoder section, there are three Transpose Convolutional Layers that perform the reverse operation of the encoder, effectively "upsampling" the compressed representation. The parameters for these layers are similar to their corresponding encoder layers but in reverse order. Lastly, the final Convolutional Layer uses a 1x2 kernel and applies a sigmoid activation function to produce the final output. The latent space obtained in the middle is used as input for a logistic regression-based classifier.

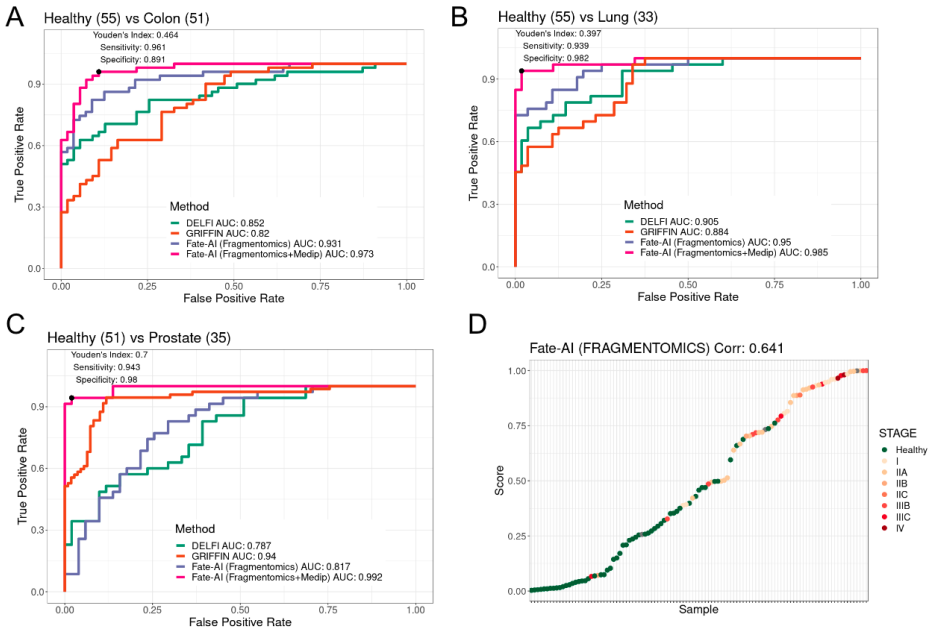


Figure 2.10. Tumor detection accuracy. A. ROC curve for the task colon cancer detection comparing Fate-AI based on Fragmentomics features and Methylation features (red curve), Fate-AI based on Fragmentomics features (purple curve), DELFI [20], and GRIFFIN [23]. B. Same as in A for Lung cancer. C. Same as in A for Prostate D. Association between Fate-AI score and clinical stage.

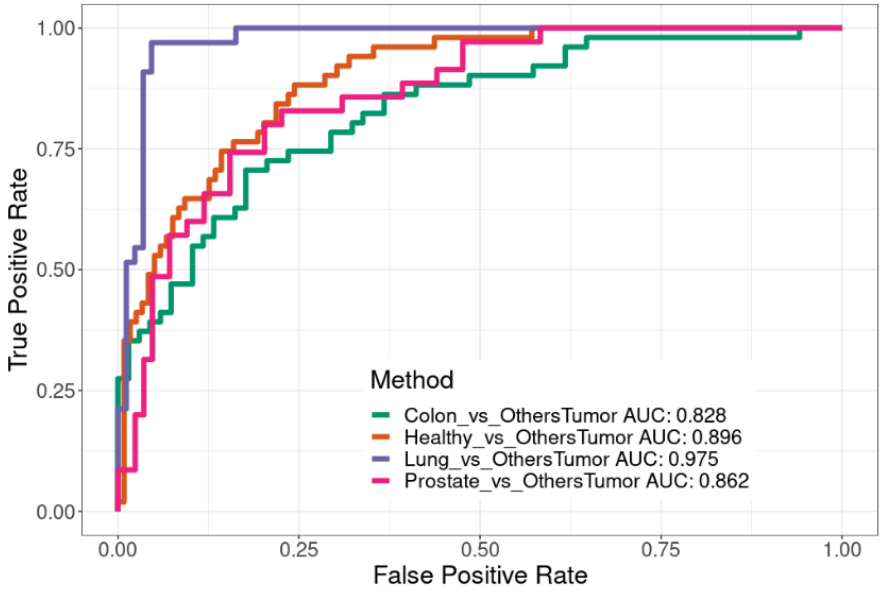


Figure 2.11. Tissue of origin. ROC curve for the task of identification of the tissue of origin with Fate-AI.

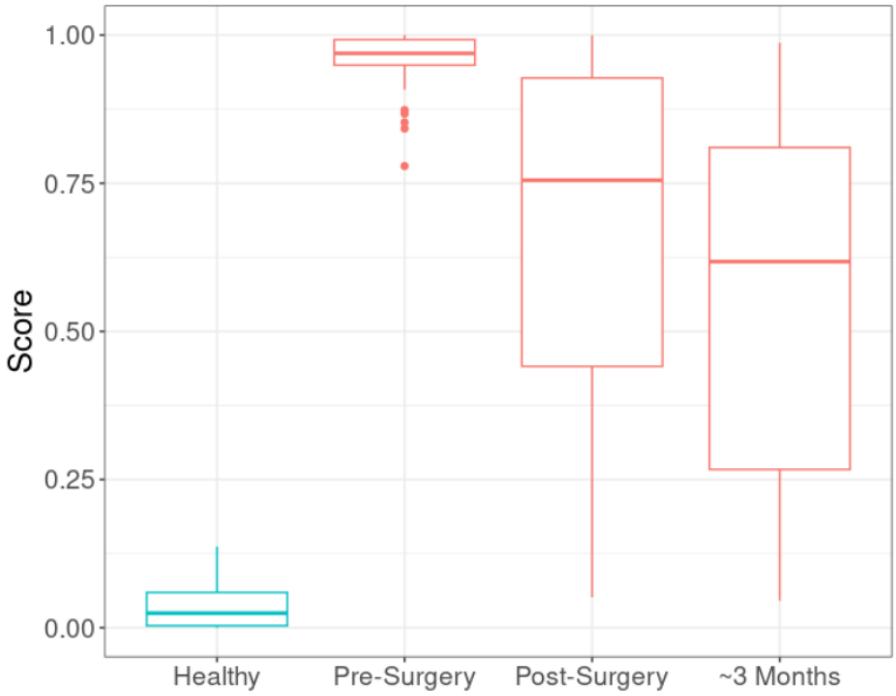


Figure 2.12. MRD Analysis of minimal residual disease (MRD). tHE prediction score of the Fate-AI model was trained only on the Healthy and pre-surgery cancer samples.

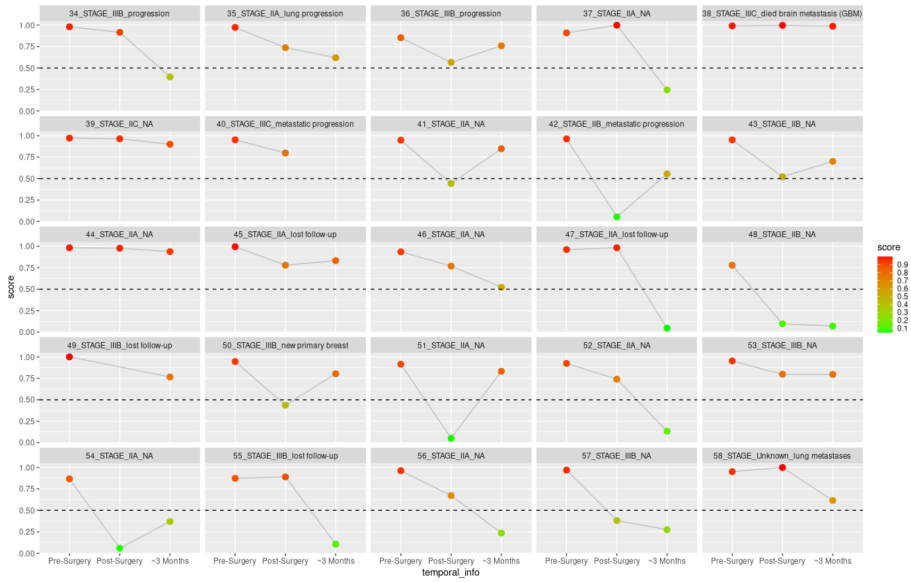


Figure 2.13. MRD clinical information The predicted score for each liquid biopsy time-point for 25 cancer samples.

Chapter 3

Prioritization of drug targets

Drug discovery is becoming more and more expensive over time despite improvements in technology. Estimates report that the number of new drugs approved per billion US dollars spent on R&D has halved roughly every nine years since 1950 [79]. The choice of appropriate therapeutic targets is one of the crucial steps in drug discovery. Machine learning approaches can exploit available high-quality and abundant data to improve decision-making in all stages of drug discovery in order to speed up the process and reduce failure rates in drug development. [101]. Here, is presented a machine-learning approach to prioritize proteins according to their similarity to approved drug targets. The main characteristic of the proposed approach is that it is completely unbiased.

The proposed method uses an extensive collection of protein features and lets the learning method score the features of approved targets. Since the aim is to extend this score to other proteins, the machine learning problem belonged to the class on positive-only problems, which can be addressed by using One Class Gaussian Processes. Also, a method for selecting the lenght-scale hyperparameter of the radial basis function kernel of the Gaussian Process is proposed. The basic idea is to use a different hyperparameter for each training sample, creating an Adaptive Kernel that varies depending on whether the training sample belongs to a sparse or dense area. The main aim is to give more importance to samples of dense areas, considered the most representative positive class samples. The validity of the proposed solution is shown in the results on the UCI benchmark

datasets, confirming that the proposed method outperforms the current state of the art based on edge-internal samples.

The development of a machine learning model based on OCGP, combined with the use of the proposed Adaptive Kernels for the hyperparameter selection, allows to define a druggability score for each protein with high performance (AUC of 0.90) on targets in clinical trials. Furthermore, several proteins outside the training and validation sets have a very high predicted score and can be considered further interesting potential candidates. The results obtained confirm the effectiveness of GPs in the one-class classification problems and that they can be improved with a correct selection of the hyperparameters. Using GP allows to get better results than and ensemble of Random Forest on the same set of features [22]. The approach has been shown to compare favorably with one-class logistic regression [89].

3.1 ML model based on OCGP

The selection and prioritization of drug targets represent a central problem in drug discovery. Drug targets are proteins associated with a particular disease process that could be addressed by a drug in order to obtain a specific therapeutic effect [100]. Experimental approaches to target identification are typically expensive, and labor intensive [7, 59]. The process from discovery to drug approval can take 10-15 years and up to several billions of investments [57]. One of the bottlenecks is the identification and prioritization of suitable drug targets. On the other hand, the increasing amount of data, which allows the creation of large scale human genomics and proteomics datasets, has the potential to reduce the work and resources needed substantially. Machine learning approaches can exploit the shared features between approved targets to select and score unknown targets [22, 36, 46, 4]. Focusing attention only on Oncology, less than 150 proteins are targets of approved drugs. These proteins can be seen as seed positive examples whose properties can be used by a learning machine to score all other potential drug targets. This kind of problem is known in machine learning as One Class Classification (OCC) or Positive Unlabeled (PU) problems [24, 15] with the additional complication of the high unbalance between the positive set and the wide set on unlabeled

samples [30]. Previous works shown that a combination of *bagging* and *easy ensemble* approaches [22, 30] can be a viable solution, which comes at the cost of the need to generate thousands of classifiers trained with samples from the unlabeled set. The proposed approach shown that the geometry of this small set of positive examples can be modeled using a class of non-parametric regressors and classifiers based on Gaussian Processes (GP) [77]. In particular, One-Class Gaussian Processes (OCGP) have been shown to outperform other kernel-based classifiers for binary and multi-class categorization of images [45, 42]. Despite the availability of a robust linear-algebra algorithm for GPs [77], the training of OCGP has some additional open questions related to the appropriate selection of hyperparameters of the kernel covariance function. Indeed, the presence of only positive samples of the training datasets makes it impossible to automatically select hyperparameters in GPs based on maximizing marginal likelihood [45].

Xiao et al. [108] recently addressed this problem. The idea is to classify the positive samples between "internal" (those in the center of the envelope containing the training set) and "edge" samples (those in the vicinity of the border of the envelope). The authors optimize the parameter by maximizing the difference between the regression function of the internal and edge samples. Other approaches use the distribution of distances among training data [51], or a different score for every positive sample based on the distances between that sample from all others [40].

Here, a novel solution is proposed for an adaptive selection of the hyperparameters of the covariance function. The proposed selection shown that a local estimate based on the distance between a sample and its neighbors can outperform the method proposed in Xiao et al. [108] both on the UCI machine learning benchmark datasets (http://homepage.tudelft.nl/n9d04/occ/index.html) and for the specific problem of drug target prioritization. In order to evaluate and compare the accuracy of prediction of the proposed method, a set of additional 277 targets for drugs in oncology clinical trials not belonging to the training set is used.

3.1.1 Protein Features

For the selection of proteins to be used as a training set, only those related to cancer drug targets were selected. Specifically, a set of approved

cancer drugs is identified based on the TTD database [52], which contained 2917 unique protein targets, of which 345 were approved, 903 clinical trials, and 1669 research targets. Only oncological targets approved and in the experimentation phase have been selected, obtaining a set of 102 approved drug targets and another 277 clinical trial targets Finally, the dataset consists of all human proteins, and each of them has 70 features obtained by combining the information in the Swiss-prot database [3], network centrality properties determined on the basis of protein-protein network information in the STRING database [95] and computationally predicting the missing data as previously described [22].

3.1.2 Gaussian Processes for OCC

Formally, a *Gaussian Process* (GP) is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution [77]. In order to specify a GP is only necessary to identify its mean and covariance functions. If the random variables represent the value of a latent function $f(\mathbf{x})$ at location $\mathbf{x}$, the mean function $m(\mathbf{x})$ and the covariance $k(\mathbf{x}, \mathbf{x}')$ of GP are:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \end{aligned} \tag{3.1}$$

and GP is thus defined as:

$$f(\mathbf{x}) \sim \mathscr{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{3.2}$$

Usually, the mean function is assumed to be zero. A GP is a very effective way to model a prior over functions simply by specifying the covariance, such as, for example, the squared exponential, which allows sampling from smooth functions. The covariance function $k(\cdot, \cdot)$ is also called the *kernel*. Given a set of, eventually noisy, training observations $\{(\mathbf{x}_i, f_i)|i = 1, \ldots, n\}$, and a set of *test* points $\{(\mathbf{x}_i^* f_i^*), |i = 1, \ldots, n'\}$, the joint distribution of the training and test output $(\mathbf{f}, \mathbf{f}^*) = (f_1, ..., f_n, f_1^*, ..., f_{n'}^*)$ is also Gaussian. GPs provide an elegant and efficient way to perform inference by incorporating the knowledge that the training data provides about the test data through *conditioning*:

$$\mathbf{f}^*|X, X^*, \mathbf{f} \sim \mathcal{N}(k(X^*, X)k(X, X)^{-1}\mathbf{f}, k(X^*, X^*) - k(X^*, X)k(X, X)^{-1}k(X, X^*)) \tag{3.3}$$

$\mathcal{N}$ is a multivariate normal distribution where the first term represents the mean vector and the second term the covariance matrix, $k(X^*, X)$ is the $n' \times n$ covariance matrix evaluated at all pairs of training and test points, analogously for the other matrices $k(X, X), k(X^*, X^*)$ and the $k(X, X^*)$. If the observations are affected by additive identically distributed Gaussian noise with variance σ_n, the $n \times n$ matrix $k(X, X)$ in equation (3.3) is replaced with $[k(X, X) + \sigma_n I]$ [77]. Other than regression, GPs can also be used for classification. In binary classification, the basic idea is to use the output of a GP regression model as a latent variable, which is then fed into a non-linear *response function*, such as the logistic or probit, that compresses the output in the range $[0, 1]$. Consider the two-class problem with target variable $y \in \{0, 1\}$. If a GP is defined over a latent variable $f(\mathbf{x})$ and then apply a *response function* $\gamma()$ which "squashes" its argument between $[0, 1]$, then a stochastic non-Gaussian process $\pi(\mathbf{x}) \stackrel{def}{=} p(y = 1|\mathbf{x}) = \gamma(f(\mathbf{x}))$ is obtained. In the case of classification, is not observed the function f but rather the input $X = \{\mathbf{x}_i | i = 1, ..., n\}$ and the corresponding class labels $y_1, ..., y_n$, and therefore is taken into consideration the value of π over the test cases $\pi(\mathbf{x}^*)$. Inference, in the case of classification, is divided into two steps:

- first computing the distribution of the latent variable corresponding to a test case:

$$p(f^*|X, \mathbf{y}, \mathbf{x}^*) = \int p(f^*|X, \mathbf{x}^*, \mathbf{f}) p(\mathbf{f}|X, \mathbf{y}) \mathrm{d}\mathbf{f} \tag{3.4}$$

  here $p(\mathbf{f}|X, \mathbf{y})$ is the posterior over the latent variables.

- the prediction is then produced averaging the response function $\gamma()$ using the distribution (3.4)

$$\bar{\pi} \stackrel{def}{=} p(y^* = 1|X, \mathbf{y}, \mathbf{x}^*) = \int \gamma(f^*) p(f^*|X, \mathbf{y}, \mathbf{x}^*) \mathrm{d}f^*. \tag{3.5}$$

Since the posterior $p(\mathbf{f}|X, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|X)$ and $p(\mathbf{y}|\mathbf{f})$ is non-Gaussian,

the integral in (3.4) cannot be analytically treated and inference is performed by a Gaussian approximation of the posterior through *Laplace Approximation* [77] or using Expectation Propagation [60]. The second one-dimensional integral (3.5) can be analytically computed in the case of probit regression or with sampling methods or analytical approximations if γ is the logistic sigmoid.

The use of GP for one class problem has been pioneered by Kemmler et al.[45]. The basic idea is to impose zero mean on the GP prior and use the value of $\mathbf{f} = 1$ in equation (3.3) on the positive examples. This will give a high probability to latent functions with values that gradually decrease for observations that are far from the positive examples. When used in combination with the choice of a smooth co-variance function, this approach results in an important subset of latent functions that can be used for OCC [45]. As shown in Figure 3.1 the predictive mean decreases for inputs far from the training data, while the predictive variance increases. The mean and variance, which are computed according to equation (3.3), both represent possible membership scores in the one-class classification problem. The predictive mean divided by the standard deviation as a combined measure to describe the uncertainty of estimation has also been proposed by Kapoor et al. [43] as an estimation of the uncertainty. Therefore, as Kemmler et al. [45] four possible scores are used for an unknown observation x^*:

- *Mean:* $\mu_* = k(x^*, X)k(X, X)^{-1}\mathbf{1}$
- *Neg. Variance:* $-\sigma_*^2 = k(x^*, X)k(X, X)^{-1}k(X, x^*) - k(x^*, x^*)$
- *Probability:* equation (3.5)
- *Heuristics:* $\mu_* \cdot \sigma_*^{-1}$

The kernel is the main component in GPs, as it represents some form of distance or similarity between data points and determines the characteristics of the function to predict. Here, the Squared Exponential (SE) kernel is used, which is the most used in GPs and thus defined:

$$k_{\mathrm{SE}}\left(x, x'\right) = \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \tag{3.6}$$

It is widely used due to its properties, which are infinitely differentiable and invariant in translation and rotation in both signal and frequency domains. It also has only a hyperparameter the length-scale (ℓ) that determines the length of the "oscillations" in the function, with a small value the function can change quickly, and conversely with large values.

3.1.3 Hyperparameter Selection

As shown in Figure 3.1, the hyperparameters significantly affect the performance of the GPs, and in particular, for OCC problems, the absence of negative samples in the training dataset does not allow automatic selection of hyperparameters through maximization of the marginal likelihood.

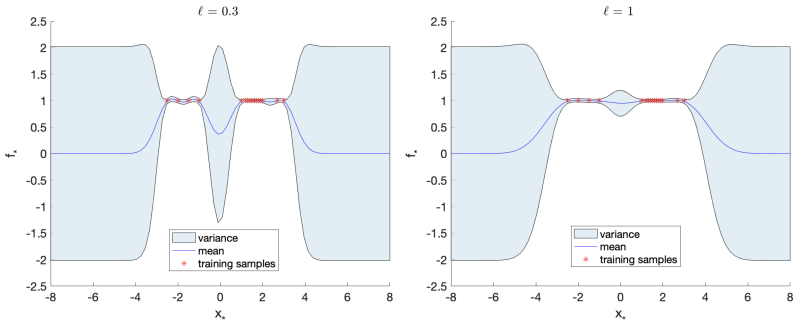


Figure 3.1. OCGP 1-D example. OCGP regression 1-D using SE kernel. The predictive distribution is visualized, mean (blue line) and variances (light blue area), and training points are marked as red asterisks. In the right panel, $\ell = 1.0$ is used, in the left panel $\ell = 0.3$.

Xiao et al. [108] propose an original solution to this problem based on the distinction between the internal samples and the edge samples of the positive class. The internal samples are assumed to be the most representative samples, instead, the edge samples that are located at the extremes of the region are considered the samples closest to the possible negative regions. Consequently, the predictions of GPs for the internal samples should be more certain, i.e. the predictive mean should be higher and the predictive variance lower, conversely for the edge samples. The authors select the optimal parameter by maximizing the Kullback-Leibler divergence between the predictions distribution of these two sets of samples.

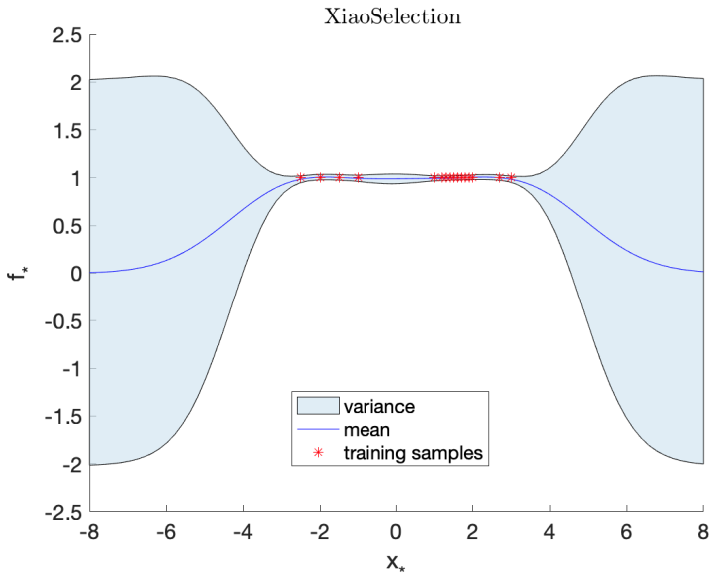


Figure 3.2. OCGP 1-D example of Xiao implementation. OCGP regression 1-D, using SE kernel and an implementation of Xiao et al. [108] hyperparameter selection method.

Li et al. [51] propose another solution to determine the hyperparameters based simply on the distribution of distances among training data. The possible hyperparameters vary between half the average of distances to nine times the average of distances and get better performance when the value is between three and seven times the average of the distances.

Kalantari et al. [40] instead propose two variants of one class of GPs; the first is OCGP-thrifty, which does not set all training target values to 1 but is based on the similarity of the training samples with the positive class. Specifically, the target value for a training sample is the average of the squared distances of that sample from all others. The second variant is OCGP-greedy, which assumes that the information from other classes is available and uses it to train a one-class model. The target training values are set as in OCGP-thrifty, also for samples of other classes. In the training phase, to find the hyperparameters, built a regression model using all the training samples. In the test phase, to calculate the predictions,

only the samples of the positive class are used. The authors show that OCGP-greedy usually obtains better results, but it cannot be applied if only samples of the positive class are labeled, as in the present case.

3.1.4 Adaptive Hyperparameter

In the adaptive kernel, no single fixed value is set for the length-scale hyperparameter of the covariance function, and this value is adapted for each training sample according to the local density. This adaptive hyperparameter depends on the local distribution of the training data, the basic idea is to give more weight to the training samples belonging to dense areas, which represent the examples sharing common features of the positive class and can be considered the most representative samples. On the other hand, less weight is given to training data lying in sparse areas, which could be less representative or outliers.

Given a sample x_i, let $\{\mathscr{N}_i^m\}_{m=1}^N$ the set of its first N neighbors ordered according to their distance from x_i. Then the value ℓ_i of each sample is set to the Euclidean distance of i-th sample from its p-th nearest neighbor.

$$\ell_i = d(x_i, \mathscr{N}_i^p) \tag{3.7}$$

Therefore, the *Adaptive Kernel* is larger in sparse areas and smaller in dense areas and is defined as:

$$k\left(x_{\mathrm{i}}, x_{\mathrm{j}}\right) = \exp\left(-\frac{\left(x_{\mathrm{i}} - x_{\mathrm{j}}\right)^2}{2\ell_{\mathrm{i}}^2}\right) \tag{3.8}$$

Since using equation (3.8) $k(x_i, x_j) \neq k(x_j, x_i)$, then the covariance matrix is symmetrized using $(K + K^T)/2$ as covariance.

Figure 3.3 shows an example in 1-D OCC setting of GP regression using zero-mean and proposed Adaptive Kernel with $p = 2$. The proposed solution allows to distinguish dense areas from areas with few samples, compared to the case where the hyperparameter is a constant value as shown in the Figure 3.1 or using the method of Xiao et al. [108] as shown in the Figure 3.2. Using the Adaptive Kernel, the predictive mean and the predictive variance tend to adapt better to the general trend of the training set. The highest test scores are obtained for test input near training

samples belonging to dense areas of the input space, which are theoretically the most representative positive samples.

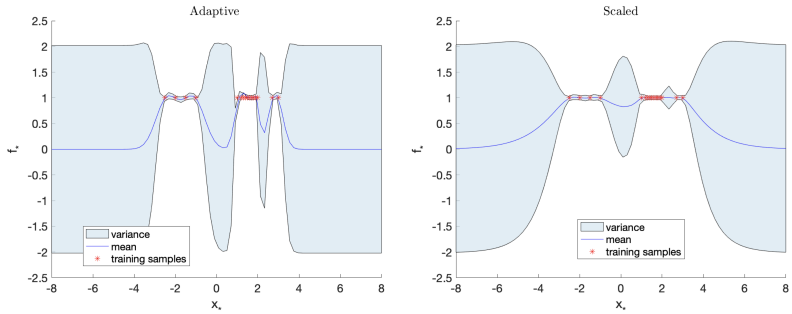


Figure 3.3. OCGP 1-D example of proposed kernel. OCGP regression 1-D, in the left panel is used the proposed Adaptive Kernel (3.8) ($p = 2$), in the right panel the Scaled Kernel (3.9) ($N = 5$).

The proposed Adaptive Kernel simply requires a search of the p-nearest neighbors of the training samples. Considering that a conventional p-nearest neighbors algorithm has $O(npd)$ complexity or $O(nd + pn)$ complexity pre-calculating and storing distances, it represents a computationally much more efficient solution, compared to the method of Xiao et al. [108] based on the edge-internal samples that have $O\left(n^3\right)$ complexity since it involves the computation of series of GPs.

In addition, another approach to automatically determine an adaptive hyperparameter: *Scaled Kernel*. This method has been successfully used in Similarity Network Fusion (SNF) [104]. In this case, hyperparameter selection combines the distance between samples and the average distance from the neighbors:

$$k\left(x_{\mathrm{i}}, x_{\mathrm{j}}\right) = \exp\left(-\frac{(x_i - x_j)^2}{\nu\varepsilon_{i,j}}\right) \tag{3.9}$$

$$\varepsilon_{i,j} = \frac{\operatorname{mean}\left(d\left(x_i, \mathscr{N}_i\right)\right) + \operatorname{mean}\left(d\left(x_j, \mathscr{N}_j\right)\right) + d\left(x_i, x_j\right)}{3} \tag{3.10}$$

In the Scaled Kernel equation (3.9), the parameter $\varepsilon_{i,j}$ combines the Euclidean distance of the samples with the average distance of samples

from the respective N nearest neighbors. Where $d(x_i, x_j)$ is the Euclidean distance, ν is a parameter that can be empirically set, and is usually set in in the range $[0.3, 0.8]$, N represents the number of neighbors considered in the calculation of the average.

Figure 3.3 also shows an example of the Scaled Kernel with $N = 5$ in the mono-dimensional space, which, like the Adaptive Kernel, allows a better distinction of the dense areas.

3.2 OCGP benchmark

In this section, before reporting the application of the adaptive OCGP to the problem of drug target prioritization, the proposed method is benchmarked against the method for hyperparameters selection for OCGP proposed in Xiao et al. [108] and also with the other one class classifiers such as support vector data description (SVDD) [97], one class SVM (OCSVM) [82] and one-class logistic regression (OCLR) [89].

3.2.1 UCI Datasets

For experiments performed on nine UCI datasets (Table 3.1) $p = 2$ is set in (3.8), while $\nu = 0.8$ and $N = 5$ were used in (3.9).

For each dataset, the class with the highest number of samples is considered as the positive class, then 80% of the positive samples are randomly chosen to build the training set while the remaining 20% of positive samples and all negative samples constitute the test set. 20 iterations of subdivision of the train and test sets are performed. The calculation of hyperparameter length-scale ℓ is performed only after normalizing data with Z-score normalization. In Table 3.2, the average results across all iterations using both predictive mean and negative variance as scores are reported. The results show that the proposed adaptive hyperparameter for both Adaptive Kernel (3.8) and Scaled Kernel(3.9) produce a significant improvement in performance when compared to the selection of the hyperparameter based on the internal and edge samples by Xiao et al. [108], in particular the Adaptive Kernel (3.8) attains the best result on the average of all datasets, with an increase from 4 to 5 percentage for the two scores mean and negative predictive variance.

dataset	features	pos	neg
Abalone	10	2770	1407
Balance	4	288	337
Biomed	5	127	67
Heart	13	164	139
Hepatitis	19	123	32
Housing	13	458	48
Ionosphere	34	225	126
Vehicle	18	647	199
Waveform	21	600	300

Table 3.1. UCI datasets

Table 3.2 also shows the results obtained using support vector data description (SVDD) [97], one class SVM (OCSVM) [82] and one-class logistic regression (OCLR) [89]. For OCSVM and OCSVM, since stationary kernels such as the rbf kernel produce the same results [80], the rbf kernel for OCSVM and a polynomial kernel of degree 3 for SVDD are used. The results confirm that Gaussian Processes are particularly suited for one-class problems, with respect to other approaches as also reported in previous works [45].

Since the proposed adaptive kernels depend on some parameters, such as p for the Adaptive Kernel and N for the Scaled Kernel, changes in performance as a function of the choice of these parameters are analyzed. The results reported below show the AUC measurement on the predictive mean obtained from the average of the 20 random splits of each dataset as a function of the parameters. In the case of the Scaled Kernel (3.9), whose results are shown in Figure 3.4, the AUC is almost constant for all of the datasets, demonstrating that this kernel is very little affected by variation of the parameter N. The Adaptive Kernel (3.8), reported in Figure 3.5, shows instead a slightly greater variation of performance for some datasets as a function of the p parameter.

	Xiao et al. [108]		Adaptive(3.8)		Scaled(3.9)		OCLR	OCSVM	SVDD
	μ_*	$-\sigma_*^2$	μ_*	$-\sigma_*^2$	μ_*	$-\sigma_*^2$		rbf	poly3
Abal.	0.7894	0.7897	0.7745	0.7428	0.7742	0.7092	**0.8760**	0.6471	0.8608
Bala.	0.8366	0.8735	0.9468	**0.9682**	0.8657	0.9402	0.5599	0.8266	0.7198
Biom.	0.8998	0.9036	0.9028	0.8960	0.9073	**0.9117**	0.9050	0.8129	0.8570
Hear.	0.8339	0.8379	0.8093	0.7925	**0.8408**	0.8135	0.5379	0.6880	0.7918
Hepa.	0.8378	**0.8379**	0.8006	0.7794	0.8242	0.7963	0.5829	0.7257	0.8055
Hous.	0.7917	0.7874	0.8677	**0.8680**	0.8107	0.8492	0.6742	0.8217	0.8374
Iono.	0.9265	0.9504	0.9550	0.9649	0.9697	**0.9712**	0.8107	0.9115	0.9341
Vehi.	0.5183	0.5714	0.7965	**0.8656**	0.6855	0.8187	0.7908	0.5601	0.5696
Wave.	0.7497	0.8004	0.7808	0.8167	0.8024	0.7998	**0.8348**	0.6160	0.5088
Aver.	0.7982	0.8169	0.8482	**0.8549**	0.8312	0.8455	0.7299	0.7344	0.7650

Table 3.2. AUC benchmark on UCI datasets.

3.2.2 Drug Target

The dataset used for the prioritization of Oncology Drug Targets consists of 20403 proteins, of which 102 are validated oncology targets used for training and 277 targets of clinical trial drugs. These last 277 proteins are used as validation set in the experiments. A total of 70 protein features related to properties derived from the sequence, protein functions, and network properties derived from the protein-protein interaction network are extracted, as previously reported [22]. The protein features in the dataset included continuous and categorical features, the latter are encoded with one-hot encoding and with frequency encoding. Some pre-processing steps are performed on the dataset, the features with a heavy-tailed distribution are log-transformed, and all features are scaled between $[0, 1]$ by min-max normalization. Furthermore, principal component analysis (PCA) is used to obtain the first principal components that allow 80% of the data variance to be retained.

First, OCGPs are compared with other OCCs. Table 3.3 reports the AUC obtained by the considered models and confirms that OCGPs outperform other classifiers.

Then, it is shown how the adaptive kernel can improve the classification accuracy. In what follows, $p = 30$ in equation(3.8) is used, while $\nu = 0.8$

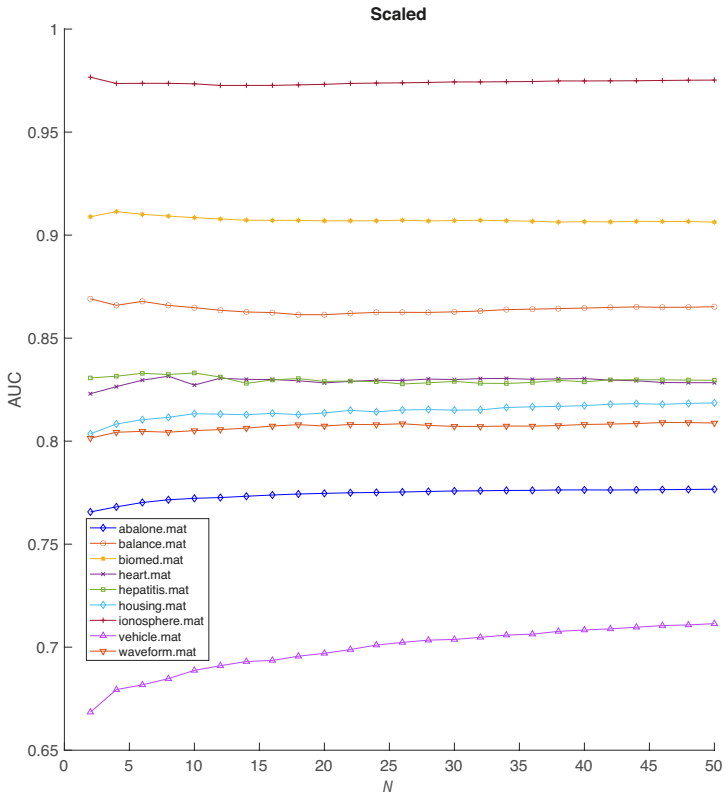


Figure 3.4. Scaled kernel benchmark. AUC scores (μ_*) on UCI datasets using the Scaled Kernel (3.9) with different values for the N parameter.

and $N = 4$ are used in equation (3.9).

In order to evaluate how the preprocessing influences the accuracy, Table 3.4 shows the results, in terms of the AUC measure on the predictive mean, obtained by adding individually the pre-processing steps described above. The results show that pre-processing leads to a significant improvement in results, and the proposed Adaptive Kernel (3.8) attains better performance than the others.

In addition, it is evaluated whether the use of a feature selection al-

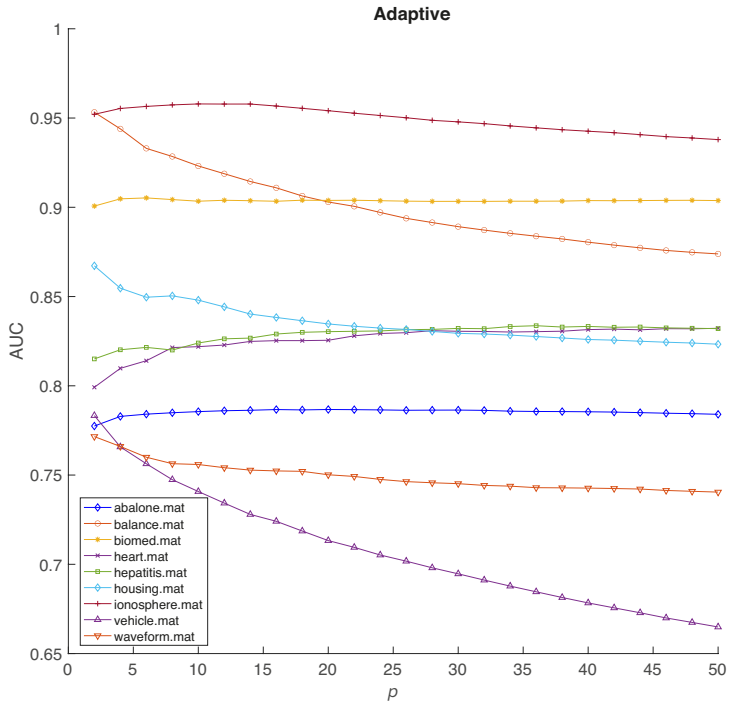


Figure 3.5. Adaptive kernel benchmark. AUC scores (μ_*) on UCI datasets using the Adaptive Kernel (3.8) with different values for the p parameter.

OC Classifiers	kernel	AUC
OCLR		0.6120
OCSVM	rbf	0.6958
SVDD	poly	0.8253
OCGP ($\ell = 0.3$)	rbf	**0.8388**

Table 3.3. Benchmark of one class classifiers on Drug Target dataset.

gorithm can possibly improve the results. Specifically, Sequential Forward Selection (SFS) [106] was used, a sequential search algorithm in which

preprocessing	Adapt. (3.8)	Xiao et al. [108]	Scaled (3.9)
scale	0.8613	**0.8680**	0.8555
scale+logtrasf.	**0.8878**	0.8610	0.8633
scale+logtrasf.+PCA	**0.8928**	0.8781	0.8759

Table 3.4. AUC comparison on the predictive mean adding each preprocessing step.

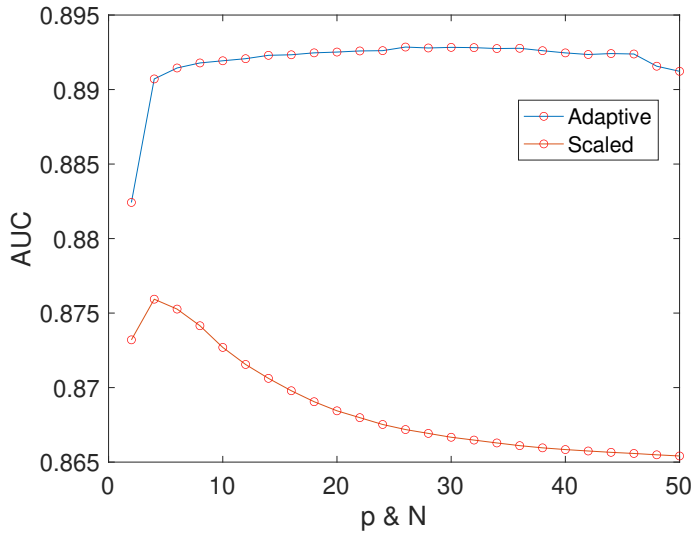


Figure 3.6. Kernels benchmark. AUC scores (μ_*) on Drug Target Dataset using Adaptive Kernel and Scaled Kernel with different values for the p and N parameters.

features are added sequentially to an empty set of candidates until the inclusion of additional features does not allow any improvement of the adopted criterion, in particular the criterion to improve is AUC measurement on the predictive mean.

The Sequential forward selection (SFS) selects 37 features and results in a significant improvement of the performance as shown in table 3.5 with Adaptive Kernel using the predictive mean as score. Interestingly, the features selected by the algorithm include network centrality measures (be-

tweenness, degree page-rank, closeness) as well as biological processes and others. Indeed, it is expected because the interaction between drugs and their targets activates signaling cascades through Protein-Protein Interaction networks, causing downstream perturbations in the cell's transcriptome. A (PPI) network thus models the cascade of relationships between targets and proteins by using physical contacts, genetic interactions, and functional relationships.

The Adaptive Kernel outperforms other methods on this dataset, but the scores obtained for test inputs differed for very low values. This is due to the hyperparameters computed before preprocessing that have high values, which consequently results in kernel values close to 0 after division with ℓ. For this reason, to obtain better dynamics, which guarantees a better interpretability of results, logarithm or square root can be applied to transform the hyperparameters computed before the preprocessing. A choice that guarantees performances comparable to the previous results, as shown in the table 3.5 where the hyperparameters for the Adaptive Kernel are log-transformed.

Figure 3.7 shows the comparison of the prediction of scores for the approved targets, clinical targets, and all other proteins. As expected, the 102 approved targets of the training set had the highest score, with a median of 0.92. Instead, for the test set, the independent set of 277 cancer clinical targets was characterized by a high median score of 0.77, unlike the rest of the proteins, which had a median score of 0.43 in the unlabeled set. Although the majority of these proteins have a lower score predicted by the proposed model, this set contains outliers with a high score that can be considered interesting potential drug targets, such, for example, the 171 outliers with scores greater than 0.91 in the boxplot of unlabeled proteins represented in red in Figure 3.7.

Some of these outliers are the subject of recent studies indicating their use as a target in oncological diseases. Among these in particular to be noted the proteins shown in table 3.7: IL7R is considered in [19] as a potential target of further therapy for leukemia patients, since the targeting of IL-7Rα signaling pathways has the potential to reduce cell proliferation and survival. JAG1 and DLL4 are the most important ligands of Notch signaling, which has a key role in the development and progression of cancer and represents an important therapeutic target, e.g. in several

Hyperparameter Selection	μ_*	$-\sigma_*^2$	Eq. (3.5)	$\mu_* \cdot \sigma_*^{-1}$
Xiao	0.8781	0.8705	0.8783	0.8773
Xiao + SFS	0.8881	**0.8717**	0.8881	0.8861
Adaptive	0.8883	0.8667	0.8878	0.8860
Adaptive + SFS	**0.9008**	0.8677	**0.9002**	**0.8981**
Scaled	0.8759	0.8500	0.8765	0.8755
Scaled + SFS	0.8911	0.8569	0.8907	0.8899

Table 3.5. AUC comparison of hyperparameter selection methods with SFS feature selection.

Hyperparameter Selection	$\min(\ell)$	$\max(\ell)$
Xiao	10.3621	
Xiao + SFS	6.7363	
Adaptive	4.1693	6.0390
Adaptive + SFS	3.9048	5.2675

Table 3.6. Range of values of selected hyperparameters.

studies the blocking of their signaling in tumors has shown interruption of angiogenesis and inhibition of tumor growth [68, 41]. PDGF and/or PDGF receptors are overexpressed or mutated in different tumors then their targeting can be beneficial in tumor treatment [31, 69], e.g. targeting PDGFRA with crenolanib has shown to significantly prevent tumor growth in inflammatory breast cancer (IBC) [38]. Moreover, Epiregulin (EREG) is identified as a possible target in lung cancer [5], particularly for Non-small-cell lung carcinoma (NSCLC) [93]. Adiponectin (ADIPOQ) is considered a potential target for many human disorders, including in particular prostate cancer [44, 34]. FGF10 is considered in several studies a possible target in particular of Pancreatic ductal adenocarcinoma (PDAC) [18, 65]. FZD2 is correlated with different cancers as shown in several studies, e.g. [35] confirms its oncogenic role in tongue cancer, and that it can be taken into account as a therapeutic target.

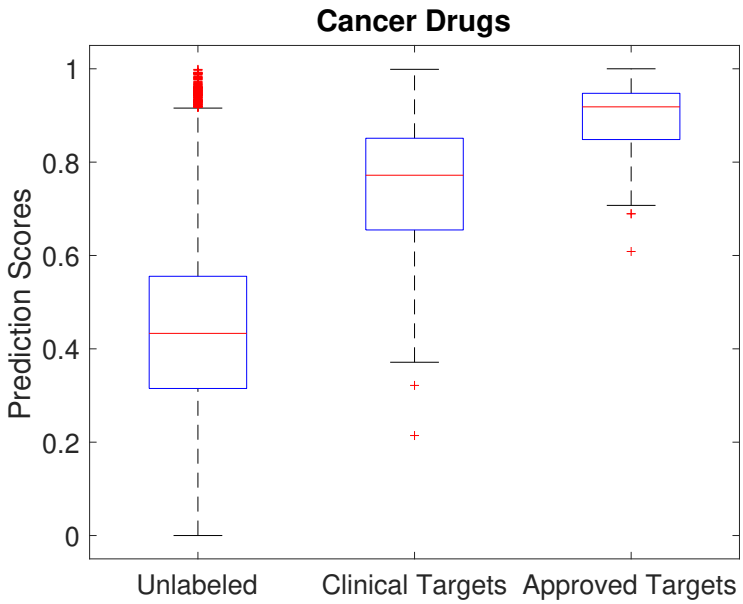


Figure 3.7. Distribution of scores. Distribution of predictions scores among the training set (approved targets), validation set (clinical trial) and the rest of the proteins. Median score: Unlabeled 0.4331, Clinical Trial 0.77196, Approved Targets: 0.9184

3.3 Findings

The proposed solution involves utilizing a large collection of protein features and One Class Gaussian Processes (GP) to score protein targets. It provides a more accurate and efficient way to identify the best protein targets for oncology drug development. Compared to an ensemble of Random Forest, which is another machine learning technique, using GP produces better and more accurate results. This is because GP algorithm models the uncertainty of the data, which is a significant advantage when dealing with biological data that can be noisy and complex.

By developing a machine learning model, it is now possible to define a druggability score for each protein with high performance. This score helps to determine the probability of a protein being a suitable target for

Gene	Score
IL7R	0.99199
JAG1	0.99122
PDGFA	0.98761
EREG	0.98334
ADIPOQ	0.98224
FGF10	0.98015
DLL4	0.97909
FZD2	0.97304

Table 3.7. Possible drug targets among the outliers.

drug development. The druggability score is a crucial step forward in drug discovery research, as it enables to focus on the most promising protein targets, thereby saving time and resources.

The application of machine learning in drug discovery is a significant breakthrough that has the potential to revolutionize the field. The use of One Class Gaussian Processes provides a more accurate and efficient way to prioritize protein targets, and the development of a druggability score offers a promising avenue for future drug development research.

Chapter 4

Conclusions

Artificial intelligence-based computational methods, therefore, represent a key resource in high-resolution tumor heterogeneity characterization, early cancer diagnosis from liquid biopsy, and drug prioritization. These applications harness the capabilities of artificial intelligence to analyze complex biological data, extract meaningful patterns, and make predictions that can significantly impact cancer research and treatment. High-Resolution Tumor Heterogeneity Characterization is the key to precision medicine. AI allows the identification of different groups of patients with cancer, that have specific subsets of genetic and molecular variations within tumors, enabling the implementation of precision medicine approaches. i.e. predicting personalized treatment strategies based on the unique characteristics of an individual's tumor. Furthermore, as shown artificial intelligence algorithms can improve sensitivity and specificity in the early identification of circulating tumor DNA from liquid biopsies. Implementation of models integrating different epigenetic features, as shown, allows for better accuracy, this is crucial for detecting tumors at earlier and more treatable stages. Furthermore, AI model predictions allow continuous monitoring of changes in liquid biopsy data over time, offering a dynamic, real-time assessment of cancer progression or response to treatment. This interactive monitoring can inform timely adjustments to treatment plans, making it possible to predict how individual patients will respond to specific cancer treatments. This personalized approach helps prioritize which drugs are most likely to be effective for a given patient, reducing the trial-and-

error aspect of treatment selection. Finally, AI algorithms can sift through vast amounts of biological data to identify novel drug targets and potential therapeutic interventions. This accelerates drug discovery processes and increases the likelihood of finding targeted therapies for specific cancer subtypes. In summary, AI-based computational methods revolutionize cancer research and treatment by providing a deeper understanding of tumor heterogeneity, enabling early cancer detection through liquid biopsy, and facilitating the prioritization of effective drugs. These advancements can potentially transform oncology practices, leading to more personalized and precise cancer care, improved patient outcomes, and ultimately contributing to the ongoing efforts to fight cancer.

Bibliography

[1] Mihaela Angelova, Bernhard Mlecnik, Angela Vasaturo, Gabriela Bindea, Tessa Fredriksen, Lucie Lafontaine, Bénédicte Buttard, Erwan Morgand, Daniela Bruni, Anne Jouret-Mourin, et al. Evolution of metastases in space and time under immune selection. *Cell*, 175(3):751–765, 2018.

[2] Samreen Anjum, Sandro Morganella, Fulvio D'Angelo, Antonio Iavarone, and Michele Ceccarelli. Vegawes: variational segmentation on whole exome sequencing for copy number detection. *BMC Bioinformatics*, 16(1):315, 2015.

[3] Boeckmann Bairoch. The swiss-prot protein sequence data bank. *Nucleic Acids Res*, 1991.

[4] Tala M Bakheet and Andrew J Doig. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4):451–457, 2009.

[5] Alison K. Bauer, Kalpana Velmurugan, Ka-Na Xiong, Carla-Maria Alexander, Julie Xiong, and Rana Brooks. Epiregulin is required for lung tumor promotion in a murine two-stage carcinogenesis model. *Molecular Carcinogenesis*, 56(1):94–105, 2017.

[6] Davide Bedognetti, Michele Ceccarelli, Lorenzo Galluzzi, Rongze Lu, Karolina Palucka, Josue Samayoa, Stefani Spranger, Sarah Warren, Kwok-Kin Wong, Elad Ziv, et al. Toward a comprehensive view of cancer immune responsiveness: a synopsis from the sitc workshop. *Journal for immunotherapy of cancer*, 7(1):1–23, 2019.

[7] Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511–516, 2019.

[8] Fenglong Bie, Zhi-Jie Wang, Yulong Li, Wei Guo, Yuanyuan Hong, Tiancheng Han, Fang Lv, Shunli Yang, Suxing Li, Xi Li, Peiyao Nie, Shun Xu, Ruochuan Zang, Moyan Zhang, Peng Song, Feiyue Feng, Jianchun Duan, Guangyu Bai, Yuan Li, and Shugeng Gao. Multimodal analysis of cell-free dna whole-methylome sequencing for cancer detection and localization. *Nature Communications*, 14, 09 2023.

[9] Mylan R Blomquist, Shannon Fortin Ensign, Fulvio D'Angelo, Joanna J Phillips, Michele Ceccarelli, Sen Peng, Rebecca F Halperin, Francesca P Caruso, Luciano Garofano, Sara A Byron, et al. Temporospatial genomic profiling in glioblastoma identifies commonly altered core pathways underlying tumor progression. *Neuro-oncology advances*, 2(1):vdaa078, 2020.

[10] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008, 04 2008.

[11] Francesca Pia Caruso, Luciano Garofano, Fulvio D'Angelo, Kai Yu, Fuchou Tang, Jinzhou Yuan, Jing Zhang, Luigi Cerulo, Stefano M Pagnotta, Davide Bedognetti, et al. A map of tumor–host interactions in glioma at single-cell resolution. *Gigascience*, 9(10):giaa109, 2020.

[12] Michele Ceccarelli. A finite markov random field approach to fast edge-preserving image recovery. *Image and Vision Computing*, 25(6):792–804, 2007.

[13] Michele Ceccarelli, Floris Barthel, Tathiane Malta, Thais Sarraf Sabedot, Sofie Salama, Bradley Murray, Olena Morozova, Yulia Newton, Amie Radenbaugh, Stefano Pagnotta, Samreen Anjum, Jiguang Wang, Ganiraju Manyam, Pietro Zoppoli, Shiyun Ling, Arjun Rao, Mia Grifford, Andrew Cherniack, Hailei Zhang, and Angeliki Pantazi. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164:550–563, 01 2016.

[14] Michele Ceccarelli, Valentina De Simone, and Almerico Murli. Well-posed anisotropic diffusion for image denoising. *IEE Proceedings-Vision, Image and Signal Processing*, 149(4):244–252, 2002.

[15] Luigi Cerulo, Charles Elkan, and Michele Ceccarelli. Learning gene regulatory networks from only positive and unlabeled data. *BMC bioinformatics*, 11(1):1–16, 2010.

[16] Ronan Chaligne, Federico Gaiti, Dana Silverbush, Joshua S Schiffman, Hannah R Weisman, Lloyd Kluegel, Simon Gritsch, Sunil D Deochand, L Nicolas Gonzalez Castro, Alyssa R Richman, et al. Epigenetic encoding,

heritability and plasticity of glioma transcriptional cell states. *Nature Genetics*, pages 1–11, 2021.

[17] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on image processing*, 6(2):298–311, 1997.

[18] Natasha Clayton and Richard Grose. Emerging roles of fibroblast growth factor 10 in cancer. *Frontiers in Genetics*, 9, 10 2018.

[19] Sarah Cramer, Peter Aplan, and Scott Durum. Therapeutic targeting of il-7r signaling pathways in all treatment. *Blood*, 128, 06 2016.

[20] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel Bruhm, Sarah Jensen, Jamie Medina, Carolyn Hruban, James White, Doreen Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian Woodward, Hatim Husain, and Victor Velculescu. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570, 06 2019.

[21] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[22] Zoltán Dezső and Michele Ceccarelli. Machine learning prediction of oncology drug targets based on protein and network properties. *BMC Bioinformatics*, 21, 12 2020.

[23] Anna-Lisa Doebley, Minjeong Ko, Hanna Liao, A. Cruikshank, Katheryn Santos, Caroline Kikawa, Joseph Hiatt, Robert Patton, Navonil De Sarkar, Katharine Collier, Anna Hoge, Katharine Chen, Anat Zimmer, Zachary Weber, Mohamed Adil, Jonathan Reichel, Paz Polak, Viktor Adalsteinsson, Peter Nelson, and Gavin Ha. A framework for clinical cancer subtyping from nucleosome profiling of cell-free dna. *Nature Communications*, 13, 12 2022.

[24] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

[25] Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J Park, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell rna-seq data. *Genome research*, 28(8):1217–1227, 2018.

[26] Véronique Frattini, Stefano M Pagnotta, Jerry J Fan, Marco V Russo, Sang Bae Lee, Luciano Garofano, Jing Zhang, Peiguo Shi, Genevieve Lewis, Heloise Sanson, et al. A metabolic function of fgfr3-tacc3 gene fusions in cancer. *Nature*, 553(7687):222–227, 2018.

[27] Ruli Gao, Shanshan Bai, Ying C Henderson, Yiyun Lin, Aislyn Schalck, Yun Yan, Tapsi Kumar, Min Hu, Emi Sei, Alexander Davis, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature biotechnology*, 39(5):599–608, 2021.

[28] Luciano Garofano, Simona Migliozzi, Young Taek Oh, Fulvio D'Angelo, Ryan D Najac, Aram Ko, Brulinda Frangaj, Francesca Pia Caruso, Kai Yu, Jinzhou Yuan, et al. Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities. *Nature cancer*, 2(2):141–156, 2021.

[29] Diana Han and Dennis Lo. The nexus of cfdna and nuclease biology. *Trends in Genetics*, 37, 05 2021.

[30] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[31] Carl-Henrik Heldin. Targeting the pdgf signaling pathway in tumor treatment. *Cell communication and signaling : CCS*, 11:97, 12 2013.

[32] Katherine Hoadley, Christina Yau, Denise Wolf, Andrew Cherniack, David Tamborero, Sam Ng, Mark Leiserson, Shubin Niu, Michael Mclellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam Margolin, Laura van 't Veer, Nuria López-Bigas, and Lihua Zou. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158, 08 2014.

[33] Charlotte Hodson, Andrew Purkiss, Jennifer Anne Miles, and Helen Walden. Structure of the human fancl ring-ube2t complex reveals determinants of cognate e3-e2 selection. *Structure*, 22(2):337–344, 2014.

[34] Xiaobo Hu, Cong Hu, Caiping Zhang, Min Zhang, Shiyin Long, and Zhaohui Cao. Role of adiponectin in prostate cancer. *International braz j urol : official journal of the Brazilian Society of Urology*, 45:220–228, 03 2019.

[35] Li Huang, Er-Ling Luo, Jing Xie, Rui-Huan Gan, Lin-Can Ding, Bo-Hua Su, Yong Zhao, Li-Song Lin, Dali Zheng, and You-Guang Lu. Fzd2 regulates cell proliferation and invasion in tongue squamous cell carcinoma. *International Journal of Biological Sciences*, 15:2330–2339, 08 2019.

[36] Zerrin Isik, Christoph Baldow, Carlo Vittorio Cannistraci, and Michael Schroeder. Drug target prioritization by perturbed gene expression and network information. *Scientific reports*, 5:17417, 2015.

[37] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 11 2019.

[38] Madhura Joglekar-Javadekar, Steven Van Laere, Michael Bourne, Manal Moalwi, Pascal Finetti, Peter Vermeulen, Daniel Birnbaum, Luc Dirix, Naoto Ueno, Monique Carter, Justin Rains, Abhijit Ramachandran, Francois Bertucci, and Kenneth van Golen. Characterization and targeting of platelet-derived growth factor receptor alpha (pdgfra) in inflammatory breast cancer (ibc). *Neoplasia*, 19:564–573, 07 2017.

[39] William G Kaelin. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698, 2005.

[40] Leila Kalantari, Paul Gader, Sarah Graves, and Stephanie Bohlman. One-class gaussian process for possibilistic classification using imaging spectroscopy. *IEEE Geoscience and Remote Sensing Letters*, 13:1–5, 06 2016.

[41] Thaned Kangsamaksin, Aino Murtomaki, Natalie M. Kofler, Henar Cuervo, Reyhaan A. Chaudhri, Ian W. Tattersall, Paul E. Rosenstiel, Carrie J. Shawber, and Jan Kitajewski. Notch decoys that selectively block dll/notch or jag/notch disrupt angiogenesis by unique mechanisms to inhibit tumor growth. *Cancer Discovery*, 5(2):182–197, 2015.

[42] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[43] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, June 2010.

[44] Hanuma Kumar Karnati, Manas Kumar Panigrahi, Yazhou Li, David Tweedie, and Nigel H Greig. Adiponectin as a potential therapeutic target for prostate cancer. *Current pharmaceutical design*, 23(28):4170—4179, 2017.

[45] Michael Kemmler, Erik Rodner, and Joachim Denzler. One-class classification with gaussian processes. pages 489–500, 11 2010.

[46] Baeksoo Kim, Jihoon Jo, Jonghyun Han, Chungoo Park, and Hyunju Lee. In silico re-identification of properties of drug target proteins. *BMC bioinformatics*, 18(7):248, 2017.

[47] James Kim and John Minna. Ap-1 leads the way in lung cancer transformation. *Developmental Cell*, 57:292–294, 02 2022.

[48] Michael Klutstein, Deborah Nejman, Razi Greenfield, and Howard Cedar. Dna methylation in cancer and aging. *Cancer Research*, 76, 06 2016.

[49] Hae-Ock Lee, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho, Valentina Pomella, Ben Van den Bosch, Jasper Vanhecke, Sara Verbandt, Hyekyung Hong, Jae-Woong Min, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics*, 52(6):594–603, 2020.

[50] Jin-Ku Lee, Jiguang Wang, Jason Sa, Erik Ladewig, Hae-Ock Lee, In-Hee Lee, Hyunju Kang, Daniel Rosenbloom, Pablo Camara, Zhaoqi Liu, Patrick Nieuwenhuizen, Sang Jung, Seung Choi, Junhyung Kim, Andrew Chen, Kyu-Tae Kim, Sang Shin, Yunjee Seo, Jin-Mi Oh, and Do-Hyun Nam. Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nature Genetics*, 49, 03 2017.

[51] Nannan Li, Xinyu Wu, Huiwen Guo, Dan Xu, Yongsheng Ou, and Yen-Lun Chen. Anomaly detection in video surveillance via gaussian process. *International Journal of Pattern Recognition and Artificial Intelligence*, 29:150426191333005, 04 2015.

[52] Yinghong Li, Chun Yu, Xiao Li, Peng Zhang, Jing Tang, Qingxia Yang, Tingting Fu, Xiaoyu Zhang, Xuejiao Cui, Gao Tu, Yang Zhang, Shuang Li, Fengyuan Yang, Qiu Sun, Chu Qin, Xian Zeng, Zhe Chen, Yu Chen, and Feng Zhu. Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic acids research*, 46, 11 2017.

[53] M Liguori, E Digifico, A Vacchini, R Avigni, FS Colombo, EM Borroni, FM Farina, S Milanesi, A Castagna, L Mannarino, et al. The soluble glycoprotein nmb (gpnmb) produced by macrophages induces cancer stemness and metastasis via cd44 and il-33. *Cellular & Molecular Immunology*, 18(3):711–722, 2021.

[54] Ruiyang Liu, Qingsong Gao, Steven Foltz, Jared Fowles, Lijun Yao, Julia Wang, Song Cao, Hua Sun, Michael Wendl, Sunantha Sethuraman, Amila

Weerasinghe, Michael Rettig, Erik Storrs, Christopher Yoon, Matthew Wyczalkowski, Joshua McMichael, Daniel Kohnen, Justin King, Scott Goldsmith, and Li Ding. Co-evolution of tumor and immune cells during progression of multiple myeloma. *Nature Communications*, 12, 05 2021.

[55] Dennis Lo, Diana Han, Jiang Peiyong, and Rossa Chiu. Epigenetics, fragmentomics, and topology of cell-free dna in liquid biopsies. *Science*, 372:eaaw3616, 04 2021.

[56] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[57] Neel S Madhukar, Prashant K Khade, Linda Huang, Kaitlyn Gayvert, Giuseppe Galletti, Martin Stogniew, Joshua E Allen, Paraskevi Giannakakou, and Olivier Elemento. A bayesian machine learning approach for drug target identification using diverse data types. *Nature communications*, 10(1):1–14, 2019.

[58] Alberto Magi, Lorenzo Tattini, Ingrid Cifola, Romina D'Aurizio, Matteo Benelli, Eleonora Mangano, Cristina Battaglia, Elena Bonora, Ants Kurg, Marco Seri, Pamela Magini, Betti Giusti, Giovanni Romeo, Tommaso Pippucci, Gianluca De Bellis, Rosanna Abbate, and Gian Franco Gensini. Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biology*, 14(10):R120, 2013.

[59] James M McFarland, Zandra V Ho, Guillaume Kugener, Joshua M Dempster, Phillip G Montgomery, Jordan G Bryan, John M Krill-Burger, Thomas M Green, Francisca Vazquez, Jesse S Boehm, et al. Improved estimation of cancer dependencies from large-scale rnai screens using model-based normalization and data integration. *Nature communications*, 9(1):1–13, 2018.

[60] Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[61] Sandro Morganella and Michele Ceccarelli. Vegamc: a r/bioconductor package for fast downstream analysis of large array comparative genomic hybridization datasets. *Bioinformatics*, 28(19):2512–2514, 2012.

[62] Sandro Morganella, Luigi Cerulo, Giuseppe Viglietto, and Michele Ceccarelli. Vega: Variational segmentation for copy number detection. *Bioinformatics*, 26(24):3020–3027, 2010.

[63] Sandro Morganella, Stefano Maria Pagnotta, and Michele Ceccarelli. Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics*, 27(21):2949–2956, 2011.

[64] David Bryant Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989.

[65] Rodrick Ndlovu, Lian-Cheng Deng, Jin Wu, Xiao-Kun Li, and Jin-San Zhang. Fibroblast growth factor 10 in pancreas development and pancreatic cancer. *Frontiers in Genetics*, 9:482, 10 2018.

[66] Cyril Neftel, Julie Laffy, Mariella G Filbin, Toshiro Hara, Marni E Shore, Gilbert J Rahme, Alyssa R Richman, Dana Silverbush, McKenzie L Shaw, Christine M Hebert, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, 178(4):835–849, 2019.

[67] Pier Nuzzo, Jacob Berchuck, Keegan Korthauer, Sándor Spisák, Amin Nassar, Sarah Abou Alaiwi, Ankur Chakravarthy, Shu Yi (Roxana) Shen, Ziad Bakouny, Boccardo Francesco, John Steinharter, Gabrielle Bouchard, Catherine Curran, Wenting Pan, Sylvan Baca, Ji-Heui Seo, Gwo-Shu Lee, M. Michaelson, Steven Chang, and Matthew Freedman. Detection of renal cell carcinoma using plasma and urine cell-free dna methylomes. *Nature Medicine*, 26, 07 2020.

[68] Chern Ein Oon, Esther Bridges, Helen Sheldon, Richard C. A. Sainson, Adrian M. Jubb, Helen Turley, Russell D. Leek, Francesca M. Buffa, Adrian L Harris, and Ji-Liang Li. Role of delta-like 4 in jagged1-induced tumour angiogenesis and tumour growth. *Oncotarget*, 8:40115 – 40131, 2017.

[69] Natalia Papadopoulos and Johan Lennartsson. The pdgf/pdgfr pathway as a drug target. *Molecular aspects of medicine*, 62, 11 2017.

[70] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[71] Jiang Peiyong, Kun Sun, Wenlei Peng, Suk Hang Cheng, Meng Ni, Philip Yeung, Macy Heung, Tingting Xie, Huimin Shang, Ze Zhou, Rebecca Chan, John Wong, Vincent Wong, Liona Poon, Tak Leung, W.K. Lam, Jason Chan, Henry Chan, KC Allen Chan, and Dennis Lo. Plasma dna end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discovery*, 10:CD–19, 02 2020.

[72] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

[73] Morgane Pierre-Jean, Guillem Rigaill, and Pierre Neuvial. Performance evaluation of dna copy number segmentation methods. *Briefings in bioinformatics*, 16(4):600–615, 2015.

[74] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[75] Sidharth Puram, Itay Tirosh, Anuraag Parikh, Anoop Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina Luo, Edmund Mroz, Kevin Emerick, Daniel Deschler, Mark Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James Rocco, William Faquin, Derrick Lin, Aviv Regev, and Bradley Bernstein. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171, 11 2017.

[76] Johannes Rainer. *EnsDb.Hsapiens.v86: Ensembl based annotation package*, 2017. R package version 2.99.0.

[77] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[78] Stefan Salcher, Gregor Sturm, Lena Horvath, Gerold Untergasser, Georgios Fotakis, Elisa Panizzolo, Agnieszka Martowicz, Georg Pall, Gabriele Gamerith, Martina Sykora, Florian Augustin, Katja Schmitz, Francesca Finotello, Dietmar Rieder, Sieghart Sopper, Dominik Wolf, Andreas Pircher, and Zlatko Trajanoski. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *bioRxiv*, 2022.

[79] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191, 2012.

[80] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, July 2001.

[81] Heidi Schwarzenbach, Dave Hoon, and Klaus Pantel. Schwarzenbach h, hoon ds, pantel kcell-free nucleic acids as biomarkers in cancer patients. nat rev cancer 11:426-437. *Nature reviews. Cancer*, 11:426–37, 06 2011.

[82] Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. Estimating support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 07 2001.

[83] Lee Serpas, Rebecca Chan, Jiang Peiyong, Meng Ni, Kun Sun, Ali Rashidfarrokhi, Chetna Soni, Vanja Sisirak, Wing-Shan Lee, Suk Hang Cheng, Wenlei Peng, KC Allen Chan, Rossa Chiu, Boris Reizis, and Dennis Lo. Dnase1l3 deletion causes aberrations in length and end-motif frequencies in plasma dna. *Proceedings of the National Academy of Sciences*, 116:201815031, 12 2018.

[84] Shikhar Sharma, Theresa Kelly, and Peter Jones. Epigenetics in cancer. *Carcinogenesis*, 31:27–36, 09 2009.

[85] Shu Yi (Roxana) Shen, Justin Burgener, Scott Bratman, and Daniel De Carvalho. Preparation of cfmedip-seq libraries for methylome profiling of plasma cell-free dna. *Nature Protocols*, 14, 08 2019.

[86] Shu Yi (Roxana) Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael Roehrl, Dianne Chadwick, Philip Zuzarte, Ayelet Borgida, Ting Wang, Tiantian Li, Olena Kis, Zhen Zhao, Anna Spreafico, Tiago Medina, Yadon Wang, David Roulois, Ilias Ettayebi, Zhuo Chen, Signy Chow, and Daniel De Carvalho. Sensitive tumour detection and classification using plasma cell-free dna methylomes. *Nature*, 563, 11 2018.

[87] 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.

[88] Matthew W Snyder, Martin Kircher, Andrew J Hill, Riza Daza, and Jay Shendure. Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164:57–68, 01 2016.

[89] Artem Sokolov, Evan O. Paull, and Joshua M. Stuart. One-class detection of cell states in tumor subtypes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 21:405–16, 2016.

[90] Kun Sun, Jiang Peiyong, KC Allen Chan, John Wong, Yvonne Cheng, Raymond Liang, Wai Chan, Edmond Ma, Stephen Chan, Suk Hang Cheng, Rebecca Chan, Yu Tong, Simon Ng, Raymond Wong, David Hui, Tse Leung, Tak Leung, Paul Lai, Rossa Chiu, and Dennis Lo. Plasma dna tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 09 2015.

[91] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22(1):163, 2021.

[92] Wei Sun, Fred A Wright, Zhengzheng Tang, Silje H Nordgard, Peter Van Loo, Tianwei Yu, Vessela N Kristensen, and Charles M Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic acids research*, 37(16):5365–5377, 2009.

[93] Noriaki Sunaga and Kyoichi Kaira. Epiregulin as a therapeutic target in non-small- cell lung cancer. *Lung Cancer: Targets and Therapy*, 6:91–98, 10 2015.

[94] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[95] Damian Szklarczyk, Annika Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda Doncheva, John Morris, Peer Bork, Lars Jensen, and Christian von Mering. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47, 11 2018.

[96] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian Tuch, Asim Siddiqui, Kaiqin Lao, and M Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6:377–82, 05 2009.

[97] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, January 2004.

[98] Alain Thierry. Circulating dna fragmentomics and cancer screening. *Cell Genomics*, 3:100242, 01 2023.

[99] Itay Tirosh, Andrew S Venteicher, Christine Hebert, Leah E Escalante, Anoop P Patel, Keren Yizhak, Jonathan M Fisher, Christopher Rodman, Christopher Mount, Mariella G Filbin, et al. Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313, 2016.

[100] David J Triggle and John B Taylor. *Comprehensive Medicinal Chemistry II*, volume 8. Elsevier, 2006.

[101] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[102] E. S. Venkatraman and Adam B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, 01 2007.

[103] Kanchan Vishnoi, Navin Viswakarma, Ajay Rana, and Basabi Rana. Transcription factors in cancer development and therapy. *Cancers*, 12:2296, 08 2020.

[104] Bo Wang, Aziz Mezlini, Feyyaz Demir, Marc Fiume, Z. Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11, 01 2014.

[105] Qi Wang, Fei Xiong, Guanhua Wu, Wenzheng Liu, Junsheng Chen, Bing Wang, and Yongjun Chen. Gene body methylation in cancer: molecular mechanisms and clinical applications. *Clinical Epigenetics*, 14, 11 2022.

[106] A. W. Whitney. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, C-20(9):1100–1103, 1971.

[107] Zuoqiao Wu, Mary Nicoll, and Robert Ingham. Ap-1 family transcription factors: a diverse family of proteins that regulate varied cellular activities in classical hodgkin lymphoma and alk+ alcl. *Experimental Hematology and Oncology*, 10, 01 2021.

[108] Y. Xiao, H. Wang, and W. Xu. Hyperparameter selection for gaussian process one-class classification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2182–2187, 2015.

[109] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 02 2015.

[110] Hui Xu, Lei Liu, Weilin Li, Duowu Zou, Jun Yu, Lifu Wang, and Chi-Chun Wong. Transcription factors in colorectal cancer: molecular mechanism and therapeutic implications. *Oncogene*, 12 2020.

[111] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4(1):1–11, 2013.

[112] Kai Yu, Yuqiong Hu, Fan Wu, Qiufang Guo, Zenghui Qian, Waner Hu, Jing Chen, Kuanyu Wang, Xiaoying Fan, Xinglong Wu, John EJ Rasko, Xiaolong Fan, Antonio Iavarone, Tao Jiang, Fuchou Tang, and Xiao-Dong Su. Surveying brain tumor heterogeneity by single-cell RNA-sequencing of multi-sector biopsies. *National Science Review*, 7(8):1306–1318, 05 2020.

[113] Zeyuan Yu, Xiangyan Jiang, Long Qin, Haixiao Deng, Jianli Wang, Wen Ren, Hongbin Li, Lei Zhao, Huanxiang Liu, Hong Yan, et al. A novel ube2t inhibitor suppresses wnt/β-catenin signaling hyperactivation and gastric cancer progression by blocking rack1 ubiquitination. *Oncogene*, 40(5):1027–1042, 2021.

[114] Jinzhou Yuan, Hanna Mendes Levitin, Veronique Frattini, Erin C Bush, Deborah M Boyett, Jorge Samanamud, Michele Ceccarelli, Athanassios Dovas, George Zanazzi, Peter Canoll, et al. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome medicine*, 10(1):1–15, 2018.

[115] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome Biology*, 18(1):174, 2017.

Author's publications

List scientific publications during the PhD:

1. A De Falco, Z Dezso, F Ceccarelli, L Cerulo, A Ciaramella, M Ceccarelli. Adaptive one-class Gaussian processes allow accurate prioritization of oncology drug targets. Bioinformatics, Volume 37, Issue 10, May 2021, Pages 1420–1427. 10.1093/bioinformatics/btaa968
2. A De Falco, F Caruso, XD Su, A Iavarone, M Ceccarelli. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. Nature Communications 14 (1), 1074 (2023). 10.1038/s41467-023-36790-9