Dynamic Causal Inference Methods for Modelling Biological Processes in Cancer



University of
South Australia

Andres Mauricio Cifuentes Bernal

Master in Mechatronics

A thesis submitted for the degree of

Doctor of Philosophy in Computer Science

November 2023

Abstract

Cancer is a dynamic and complex system at the genetic level. Throughout the development of cancer, genetic aberrations alter the regulation of various biological processes. One of the most ambitious objectives of cancer research is to understand the dynamical behaviour of gene regulation patterns that occur during cancer development. The results of pursuing this aim will enhance our understanding of the cellular changes that occur during the evolution of the disease.

Expanding our understanding of cancer dynamics involves delving into the regulatory relationships that alter the biological processes influenced by genetic abnormalities. It also entails comprehending the dynamics of the gene interactions and collaborations that drive the progression of the disease. Modelling the dynamics of genes in cancer is a difficult task because tissue samples are usually destroyed while information is captured at one or more omics levels. Consequently, the vast majority of data gathered for cancer research are static representations of cellular states. Hence, most bioinformatics and statistical approaches in cancer research consist of static analyses of cancer data.

In most cases, explanations of the temporal patterns of gene expression (i.e. patterns in the temporal sequences) exhibited by genes during cancer progression are still inadequate. Although most current bioinformatics and statistical analyses have provided invaluable information on cancer, they overlook some significant aspects of the disease. For example, the majority of cancer-related gene regulatory networks accepted by the global scientific community correspond to comparisons between networks detected using healthy tissue samples and networks detected using cancer cells. In practice, gene regulatory networks that characterise the behaviour of the biological processes altered by cancer may not correspond to either of these two extremes. Similarly, static analyses employed to detect the genes that drive cancer do not account

for the dynamical aspect of the disease, which reduces the range of drivers that can be discovered from the data. Furthermore, static analyses overlook the dynamics of the interactions and collaborations between cancer driver genes.

This thesis expands knowledge about the dynamics of genes in cancer through novel approaches that account for the temporal aspect of gene regulations. Specifically, three fundamental aspects of the dynamics of the disease are addressed: (i) the genetic regulations that characterise altered processes that occur *during* cancer progression, (ii) the genes that drive altered biological processes in cancer, and (iii) the dynamics of the interactions and collaborations between genes that are jointly responsible for the progression of cancer.

The first approach focuses on the detection of miRNA-mRNA relationships that characterise altered biological processes that occur *during* cancer progression. The proposed method not only significantly outperforms traditional miRNA-mRNA relationship detection methods, but it also identifies miRNAs relevant to cancer. The latter is a difficult task because miRNAs are small bio-molecules that do not encode proteins, and this characteristic hinders the analysis of their functional impact in cancer.

The second approach utilises the changes in gene expression patterns that manifest throughout the development of cancer to identify genes that drive the alteration of biological processes. This approach is useful for expanding current highly confident cancer driver catalogues, as demonstrated by its ability to detect cancer drivers regardless of their mutational status. This is possible because of the incorporation of the temporal aspect of cancer, unlike traditional methods that use mutational events (usually in protein-coding genes) to define cancer drivers. This approach's capability of detecting non-mutated drivers of cancer is a significant contribution to the field because such drivers have not been extensively studied, even though recent research has shown that other aberrations besides mutations can also drive cancer.

Finally, the third approach models the interactions and collaborations between genes that jointly drive cancer progression. Experimental tests

show that this approach is capable of systematically identifying groups of genes that jointly drive cancer. Furthermore, this approach identifies groups of cancer drivers and models the dynamics of their interactions and collaborations during cancer development. It is also capable of detecting mutated and non-mutated genes that drive cancer. Thus, this approach contributes to two currently under-explored areas: the detection of genes that jointly drive cancer progression and the modelling of the dynamics of such interactions and collaborations.

These approaches can be used as complements in wet laboratory experiments. The contributions of this thesis can be translated into a better understanding of the temporal behaviour of genes throughout cancer progression, which can potentially facilitate early cancer detection and the development of new treatments.

Statement of Originality

The thesis presents the research conducted by myself and does not include any material previously submitted for a university degree or diploma, except where proper citation is provided. To the best of my knowledge, it does not contain any information previously published or created by another person, except in cases where appropriate references are given in the text. All significant contributions from others to the research presented, including those from collaborative publications, are appropriately acknowledged.

Andres Cifuentes

Digitally signed by Andres Cifuentes
Date: 2023.11.30 14:38:39 +10'30'

Andres Mauricio Cifuentes Bernal
November 2023

Acknowledgements

I am deeply grateful to many people for their support in my PhD journey, and I want to take this opportunity to express my appreciation.

First and foremost, I extend my heartfelt thanks to my supervisor, associate professor *Thuc Duy Le*. His guidance has been more than just academic; it has been a beacon of light in moments of doubt and confusion. I deeply appreciate his patience, expertise, and encouraging words. His support has not only shaped my research but have also inspired me to strive for excellence. I am greateful for his willingness to share his knowledge and his constant support that have been instrumental in making this journey a successful and enriching experience.

I also owe a great deal of gratitude to my co-supervisors, Professor *Lin Liu* and Professor *Jiuyong Li*. Their perspectives and constructive feedback have immensely contributed to the depth and quality of my research. The discussions we had and the insights they provided have been invaluable, helping me to see my work from different angles and pushing me to refine my ideas further.

To my peers, colleagues, and friends, Dr *Xiaomei Li*, and *Dr Vu Viet Hoang Pham*, I express my sincere thanks for all of their help supporting my research. Our shared experiences, from brainstorming sessions to tackling complex problems, have not only enriched my research but have also made this academic journey enjoyable and memorable. Their support, camaraderie, and collaborative spirit have been a constant source of motivation and learning.

My sincere thanks also go to in the Data Analytics group in the University of South Australia, particularly to Associate Prof. *Jixue Liu*, Dr *Debo Cheng*, Dr. *Ha Xuan Tra*n, Dr. *Sha Lu*, *Ziqi Xu*, *Liang Zhao*, *Ji-Young Park*, and *Oscar Blessed Deho* for their contributions to the discussion during the group meetings. Their support, whether it was through practical

help or guidance, has played a significant role in my academic journey. Their assistance and encouragement have been greatly appreciated.

A special acknowledgement and warm gratitude goes to my friends, Dr *Menghao Jia*, and *Hongshou Huang*. Their unwavering support, understanding, and encouragement have been a pillar of strength throughout this journey. Whether it was offering a listening ear during challenging times or celebrating small victories along the way, their friendship has been a treasured part of this experience.

I acknowledge the support of the Australian Technology Network (ATN) LATAM Research Scholarship and the Australian Research Training Program. The ATN LATAM Research Scholarship has not only provided me with the financial resources necessary to pursue my doctoral studies but has also opened doors to invaluable opportunities for academic growth and cross-cultural exchange. Likewise, the funding received from the Australian Research Training Program has been essential in supporting my journey through the PhD program. Both programs have contributed to my ability to focus on my studies and research without the burden of financial constraints.

I cannot express enough gratitude to my family – my parents *Argemiro Cifuentes Osorio* and *Lilia Bernal Pedraza*, my sister *Carol Milena Cifuentes Bernal*, as well as my nieces *Dana Valentina* and *Natalia* -. Their love, sacrifice, and belief in my abilities have been the driving force behind my perseverance. They have been my support system, providing encouragement and motivation at every step. This accomplishment is a reflection of their unwavering faith in me and their endless support, and this achievement is as much theirs as it is mine.

In closing, I wish to extend my appreciation to all those who have been a part of this journey, directly or indirectly. This thesis not only represents my academic efforts but also the collective support and encouragement of everyone mentioned here, and many others unmentioned but equally appreciated.

Andres Mauricio Cifuentes Bernal
November 2023

List of publications

1. **Andres M Cifuentes-Bernal**, Vu Vh Pham, Xiaomei Li, Lin Liu, Jiuyong Li, Thuc Duy Le, A pseudotemporal causality approach to identifying miRNA–mRNA interactions during biological processes, Bioinformatics, Volume 37, Issue 6, March 2021, Pages 807–814, https://doi.org/10.1093/bioinformatics/btaa899
2. **Andres M Cifuentes-Bernal**, Vu V H Pham, Xiaomei Li, Lin Liu, Jiuyong Li, Thuc Duy Le, Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories, Briefings in Functional Genomics, Volume 21, Issue 6, November 2022, Pages 455–465, https://doi.org/10.1093/bfgp/elac030
3. **Andres M Cifuentes-Bernal**, Lin Liu, Jiuyong Li, Thuc Duy Le Identifying cooperative genes causing cancer progression with dynamic causal inference, 2023, bioRxiv 2023.11.22.568367, https://doi.org/10.1101/2023.11.22.568367

Contents

1 Introduction **1**

1.1 Background and motivation 2

1.2 Objective and research problems 4

1.3 Contributions 4

1.4 Thesis structure 9

2 Background **10**

2.1 Gene expression and biological processes 11

2.1.1 Non-coding RNA roles in the regulation of biological processes 13

2.1.2 Dysregulation in cancer 15

2.2 Dynamical systems and stochastic processes 17

2.3 Causal inference from temporal data 20

2.4 Cancer data and data transformation 22

2.4.1 Data for cancer research 22

2.4.2 Pseudotime data 24

3 Literature review **28**

3.1 Methods for inferring gene regulatory networks 29

3.1.1 Co-expression methods 31

3.1.2 Machine learning methods 33

3.1.3 Differential equation-based methods 35

3.1.4 Causality based methods 37

3.1.4.1 Granger causality 38

3.1.4.2 Bayesian network based methods 40

3.1.4.3 Other causal methods 43

3.2 Methods for discovering cancer drivers 45

3.2.1 Mutation-based methods 47

3.2.1.1 Frequency-based methods . . . 47
3.2.1.2 Hotspot-based methods . . . 48
3.2.1.3 Functional impact-based methods . . . 50
3.2.2 Network-based methods . . . 51
3.2.3 Causality based identification of cancer drivers . . . 54
3.3 Gene Group Analysis of Cancer Drivers . . . 57

4 A pseudotemporal causality approach to identifying miRNA-mRNA interactions during biological processes 62
4.1 miRNA-mRNA regulatory relationships . . . 63
4.1.1 Invariance property of causal structures for causal inference from sequential data . . . 65
4.2 Methods . . . 68
4.2.1 Problem definition and method outline . . . 68
4.2.2 Pseudotime analysis . . . 69
4.2.3 Gene selection . . . 71
4.2.4 Identification of miRNA-mRNA regulatory relationships . . . 71
4.2.4.1 Testing violation of the invariance property . . . 72
4.2.4.2 Identifying miRNAs regulating a mRNA . . . 73
4.3 Results and discussion . . . 74
4.3.1 Identified miRNA-mRNA interactions in EMT with single-cell data . . . 74
4.3.2 Experimental confirmations of interactions discovered by *PTC* 76
4.3.3 Identification of interactions relevant to EMT using *PTC* . . . 77
4.3.4 Stability of *PTC* at different pseudotimes . . . 78
4.3.5 Performance of *PTC* relative to benchmark methods for bulk gene expression data . . . 79
4.4 Conclusion . . . 82

5 Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories 84
5.1 Introduction . . . 85
5.2 Methods . . . 89
5.2.1 Problem definition and method overview . . . 89
5.2.2 Phase 1: Network-based selection of putative driver genes . . . 91
5.2.3 Phase 2: Pseudotime series modelling and event detection . . . 92
5.2.3.1 Pseudotime series . . . 92

5.2.3.2 Event detection . . . 92
5.2.4 Phase 3: Cancer driver identification based on causal impact . 94
5.2.4.1 Counterfactual model . . . 95
5.2.4.2 Discovery of dynamic cancer drivers . . . 96
5.3 Results . . . 96
5.3.1 Dynamic cancer drivers identified from single-cell data . . . 96
5.3.2 Contribution of proposed method to expanding current cancer driver catalogues . . . 98
5.3.2.1 Stratification of inferred driver genes based on literature support . . . 100
5.3.3 Enrichment analysis of dynamic cancer drivers inferred from single-cell data . . . 101
5.3.4 Discovery of cancer drivers from bulk data using dynamical information . . . 103
5.3.5 Gene regulatory networks for dynamic cancer driver genes . 105
5.3.6 Cancer drivers suitable for biological process covariates . . . 106
5.3.7 Differentially expressed gene (DEG) analysis of DCD genes inferred from bulk data . . . 108
5.3.8 DCD approach is useful for detecting drivers from different pseudotime methods . . . 109
5.4 Discussion and conclusions . . . 110

6 Identifying cooperative genes that cause cancer progression using dynamic causal inference 113
6.1 Introduction . . . 114
6.2 Methods . . . 117
6.2.1 Problem definition . . . 117
6.2.1.1 Assumptions . . . 117
6.2.1.2 Invariance of structure of dynamic systems for causal discovery of cancer driver genes . . . 118
6.2.2 Method overview . . . 119
6.2.3 Formulation of putative models . . . 120
6.2.4 Pseudotime conversion . . . 123
6.2.5 Discovery of dynamic cancer drivers . . . 124
6.3 Experiments and results . . . 126
6.3.1 Cancer drivers discovered from TCGA breast cancer data . . . 126

6.3.1.1 Systematic detection of cancer drivers by the proposed method . . . 128
6.3.2 Identification of mutated and non-mutated cancer drivers by the proposed method . . . 130
6.3.3 Inferred cancer driver sets from single-cell data . . . 132
6.4 Conclusion . . . 137

7 Conclusions and future work **138**
7.1 Summary of contributions . . . 139
7.2 Future directions . . . 142

A Supplementary material of the PTC method **147**
A.1 Single cell data processing . . . 148
A.2 "GO biological processes" and "KEGG pathways" analysis . . . 149

B Supplementary material of the Dynamic Cancer Drivers method **150**
B.1 GO biological processes analysis . . . 151
B.2 Gene-disease association analysis (DisGeNet) . . . 152
B.3 Top 100 dynamic cancer drivers from HER2time(SC) . . . 153
B.4 Top 100 dynamic cancer drivers from VIMtime(SC) . . . 156
B.5 Stratification of inferred driver genes (intersection set) based on literature support . . . 159
B.6 Top 10 GO terms HER2time(SC) . . . 167
B.7 Top 10 GO terms VIMtime(SC) . . . 169
B.8 Top 10 Gene-Disease Associated terms (GDAs) - HER2Time(SC) . . . 170
B.9 Top 10 Gene-Disease Associated terms (GDAs) - VIMTime(SC) . . . 171
B.10 Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA. . . . 171
B.11 Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA. . . . 175
B.12 Regulatory relationships among Dynamic Cancer Drivers - HER2time(SC) 178
B.13 Regulatory relationships among Dynamic Cancer Drivers - HER2time(Bulk) 181
B.14 Not differentially expressed DCD inferred from HER2Time(Bulk) . . . 183
B.15 Not differentially expressed DCD inferred from VIMTime(Bulk) . . . 185
B.16 DCD inferred from single cell data when using Monocle3 pseudotime (event detection using HER2). . . . 187

B.17 DCD inferred from single cell data when using Monocle3 pseudotime(event detection using VIM). 190

C Supplementary material of identifying genes that cause cancer progression using dynamic causal inference **195**

C.1 Hypergeometric test for the 39 analysed targets in Bulk Data 196

D Statements of authorship **198**

D.1 A pseudotemporal causality approach to indentifying miRNA-mRNA interactions during biological processess 199

D.2 Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories 203

D.3 Identifying cooperative genes that cause cancer progression using dynamic causal inference 207

Bibliography **211**

List of Figures

2.1 Central dogma of molecular biology 12
2.2 Example of two equivalent causal models 21
2.3 Illustration of pseudotemporal analysis 26

4.1 Summary of the proposed method, *pseudotemporal causality (PTC)* . . 70
4.2 miRNA-mRNA regulatory relationships discovered by PTC. 75
4.3 Gene expression of the pair (hsa-miR-16-5p - VAMP8) as a function of the VIM-Time. 78
4.4 Confirmed interactions found by PTC in bulk data 81
4.5 Intersection of the confirmed miRNA-mRNA interactions predicted to be in the top 200 by various methods when using bulk data 82

5.1 Graphical representation of the overlap between different categories of cancer drivers 86
5.2 Summary of the *dynamic cancer driver* inference method 91
5.3 Chow-Ruskey Venn diagram of discoveries of proposed method when exploring cancer drivers following the "VIMtime(SC)" and "HER2time(SC)" *pseudotime* orders. 98
5.4 Comparison between number of novel CGC genes inferred by the proposed method and those inferred by each of the cancer driver catalogues analysed. 99
5.5 Comparison of different cancer driver inference methods with bulk data 104
5.6 Regulatory relationships among DCDs 106
5.7 Performance of the DCD method with ESR1 as a path covariate against various cancer driver inference methods for bulk data. 107

6.1 Summary of the proposed method for identifying cooperative genes that cause cancer progression 120

6.2 Inferred gene cooperation networks when using the top 50 models of the experiments with PTEN and AKT1 as target variables. 135

B.1 Top 10 Go biological terms 2021 (ranked by p-value) from our enrichment analysis for the gene set obtained from ″VIMtime(SC)″ (upper) and ″HER2time(SC)″ (lower). 151

B.2 Top 10 GDAs terms (ranked by p-value) from our enrichment analysis for the gene set obtained from ″VIMtime(SC)″ (upper) and ″HER2time(SC)″ (lower) retrieved from DisGeNET. 152

List of Tables

2.1 Common types of molecular omics data in cancer research 23

4.1 Summary of performance results for *PTC*, Pearson, and Lasso. 76
4.2 Summary of bulk data performance of *PTC* 80

5.1 Summary of inferred dynamic cancer drivers from single-cell data when using HER2 and VIM as process covariates 97
5.2 Top 5 GO biological processes (left) and DisGeNET DGAs (right) enriched terms among cancer drivers inferred from VIMtime(SC) . . 102
5.3 Summary of inferred sets of **dynamic cancer drivers** from single-cell data when using ESR1 as the process covariate 107
5.4 Comparison between DEGs and DCDs from the TCGA-BRCA dataset 109

6.1 List of the 40 Mut-driver genes used in the experiments as target variables 127
6.2 Summary of the 11 target genes with a significant overlap with the CGC gene set 128
6.3 Summary of the seven additional targets with a statistically significant overlap with the CGC gene list when using the top 150 genes ranked by the proposed method 129
6.4 Top 10 most mutated (SNV) and all non-mutated inferred cancer drivers in the intersection set of the results for the 11 target genes shown in Table 6.2 131
6.5 Summary of single-cell data experiments for the 14 target genes with a significant overlap with the CGC gene set when considering the top 80 inferred genes per target 133
6.6 Cancer driver genes inferred from the single-cell data that were detected by the proposed method in at least 10 different experiments (different target genes) 134

6.7 Enriched biological pathways for the IL1RN gene cluster (community 1 in Fig. 6.2a) in Reactome 2022, KEGG human 2021, and Elsevier Pathway Collection 136

A.1 PTC inferred genes significantly associated with KEGG pathways and GO biological processes 149

B.1 Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact. 153
B.2 Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact. 156
B.3 Stratification of inferred driver genes (intersection set) based on literature support. DCD are ranked by following Relative Causal Impact (absolute value) 160
B.4 Top 10 GO terms HER2time(SC) ranked by p-value. 167
B.5 Top 10 GO terms VIMtime(SC) ranked by p-value. 169
B.6 Top 10 GDAs HER2Time(SC) ranked by p-value 170
B.7 Top 10 GDAs VIMTime(SC) ranked by p-value 171
B.8 Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA (descending order). 172
B.9 Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA (descending order). 175
B.10 Regulatory relationships Dynamic among Cancer Drivers HER2time(SC).178
B.11 Regulatory relationships Dynamic among Cancer Drivers HER2time(Bulk).181
B.12 DCD inferred from HER2Time(Bulk) that are not differentially expressed 183
B.13 DCD inferred from VIMTime(Bulk) that are not differentially expressed 185
B.14 dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2). 187
B.15 dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM). 191

C.1 Results of hypergeometric test for the 39 analysed targets in the TCGA-BRCA dataset. 196

Chapter 1

Introduction

Cancer is a genetic disease that has become the primary cause of premature death in a significant number of countries (112 out of 183 by 2019) [1]. According to estimates from the World Health Organization, cancer was the cause of one in six deaths in 2020, with about 18 million new cases detected in the same year [1, 2]. A critical fact is that treating cancer in its early stages drastically increases the survival expectations. Therefore, to enhance a patient's survival prognosis, developing early cancer detection methods at the molecular level is crucial [3]. As the initial signs of cancer are manifested through genetic alterations, the development of such methods necessitates a comprehensive understanding of the temporal patterns of gene regulation that contribute to the onset and progression of the disease.

1.1 Background and motivation

The fact that cancer is a genetic disease that involves dynamic changes in the genome has been known for several decades. Evidence suggests that cancer progression is initiated by genetic alterations that drive multi-step changes in cells [4]. Those alterations provide cancer cells with a critical set of functional capabilities, which are commonly known as the *hallmarks of cancer* [4, 5, 6]. These hallmarks grant cancer cells an evolutionary advantage by enabling them to resist apoptosis (programmed cell death), reprogram their cellular metabolism, and avoid immune destruction, among others [6].

The development of such functional capacities in a cell is a result of dysregulation in gene expression. Gene expression involves the conversion of genetic information encoded in the genes into functional products, such as RNA and proteins. Gene expression is a highly regulated dynamical process that plays a critical role in determining the state and functions of a cell at a given time. Disturbances in such a highly regulated process lead to various diseases, including cancer [7]. Hence, understanding and modelling the dynamics of gene expression *during* cancer progression is a critical endeavour in cancer research.

However, the gene dynamics and gene regulation patterns that occur during cancer progression are poorly understood. For years, most biological analyses of cancer at the genetic level have favoured Darwinian models (i.e. natural selection approach) over dynamic models of genes. Although Darwinian models accurately describe the dynamics of cancer cell populations [8, 9], this ecological approach is

insufficient for a complete understanding of cancer dynamics. A full understanding additionally requires accounting for the dynamics of gene expression patterns.

Because cancer evolves progressively, models of cancer (and by extension, bioinformatics methods used to study it) should include the temporal aspect of the disease. Changes at the cellular level caused by the dysregulation that occurs during cancer development are progressive. Such changes constitute a continuous process, transforming cells from normal to cancerous [3]. Therefore, a comprehensive model of cancer should model the changes that occur throughout cancer progression. This is not a simple task because the screening and data collection of cancer in humans usually requires phenotypical manifestations of the disease [10, 11, 12]. Such phenotypical manifestations occur in relatively advanced stages of cancer; thus, data on tumour initiation is usually incomplete.

Not only is there a lack of data on the early stages of cancer, but the vast majority of currently available cancer data are not suitable for use in dynamic modelling. Most cancer data are recorded during the examination of tissue samples (biopsies) taken from patients with the disease. A common practice is to simultaneously take samples of cancerous and normal tissue from the same patient. This enables a comparison between normal and cancerous conditions. Accordingly, a typical assumption is that the differences between cancerous and normal tissues reflect the changes that characterise the disease. However, there are several problems with this type of analysis, and they limit its usability.

The first limitation is that the temporal aspect of the disease is not accounted for in the data. Data from tissue biopsies are considered static because they are recorded at a single instant in time for each patient. Therefore, direct analyses of this type of data are categorised as static studies. Although static studies can provide important insights into the disease, they yield minimal information about its dynamics.

The second limitation is the fact that genetic cancer research is strongly biased towards mutational events that transform genes into cancer drivers. This perspective neglects the fact that cancer itself is a megabiological process. Specifically, cancer is a collection of altered biological processes. A biological process is a series of sequential events that allow an organism to accomplish a genetically programmed task (e.g. cell division). Altered biological processes confer on cancer cells the

aforementioned *hallmarks of cancer*. However, studies focusing on the contrast of mutational states between normal and cancerous conditions often overlook the role of biological processes in cancer progression. Therefore, this type of analysis ignores critical aspects of the disease.

1.2 Objective and research problems

The overall objective of this thesis is to investigate the underlying dynamics of cancer, with a particular focus on the genes involved in cancer genesis and progression. Specifically, three of the most relevant sub-problems related to our current lack of knowledge in this context are addressed:

- **Research Question 1**: What gene regulatory relationships characterise the dynamics of biological processes during the progression of cancer?

- **Research Question 2**: What is the causal effect of genes on the biological processes responsible for the genesis and development of cancer?

- **Research Question 3**: What are the dynamics of gene collaborations and interactions that contribute to the progression of cancer?

1.3 Contributions

This thesis focuses on the key fact that cancer is a conglomeration of dysregulated biological processes that progress temporally. The main hypothesis is that accounting for the temporal aspect of biological processes enhances the results that can be obtained from data-driven cancer analyses. Consequently, in the research conducted for this thesis, three computational methods that enable cancer data analyses to account for the underlying dynamic information retrieved from the data are developed. Each method addresses one of the three problems stated above.

The first problem is related to the regulation of biological processes in cancer progression. A biological process originates from the interaction between elementary units such as genes. In these interactions, certain genes influence changes in other genes, causing the biological process to materialise. This influence is commonly

called "regulation". Together, these interactions comprise what are known as gene regulatory networks (GRNs).

Even though the inference of GRNs is not recent (see section 2.1 for a review), the most accepted representations of regulatory networks originate from static data. In the case of cancer, certain regulatory network representations are inferred from normal tissue and others are inferred from cancerous tissue. However, in general, neither of these two representations accurately reflect the regulatory relationships that exist between genes throughout the progression of cancer. For this reason, this thesis aims to identify the *regulatory relationships that characterise biological processes throughout cancer progression* rather than the differential GRNs between cancerous and normal tissues.

In addition, traditional GRN inference methods usually rely on association measurements such as Pearson/Spearman/Kendall coefficients, Euclidean distances, or mutual information. However, regulatory relationships are causal, and thus the causal nature of the relationship needs to be assessed.

The method developed in this thesis (explained in detail in chapter 3) utilises the expression of micro RNAs (miRNAs) and messenger RNAs (mRNAs), two elemental genetic entities, to infer the miRNA-mRNA regulatory networks that characterise biological processes during cancer progression.

The first method addresses two key emerging challenges in characterising a biological process. First, it requires data to be collected at different points in time during the process. Second, testing causality during the biological process is crucial for miRNA-mRNA relationships, which are considered causal. While a substantial body of research that utilises gene expression profiles exists, most of it examines static data corresponding to different biological states (e.g. normal vs cancerous) rather than the continuous transition that occurs in biological processes.

Several widely accepted methods only identify the correlation or association between miRNA and mRNA expression levels. Other methods infer the causality, but at the cost of high computational complexity. Based on the key idea that miRNA-mRNA relationships that characterise a biological process have a causal link throughout the process, a method that aims to find miRNA-mRNA causal

relationships by testing a property of causal systems known as *invariance* [13] is designed in this thesis.

In the experiments that tested this method, the process selected was the *epithelial to mesenchymal transition* (EMT), a biological process strongly related to cancer progression and metastasis. The method was applied to a single-cell dataset (K562 cells) and the results are validated by comparing the method's inferences against a database containing experimentally confirmed miRNA-mRNA relationships. Because most methods for inferring gene regulatory relationships use bulk data, benchmark experiments with bulk data were established.

The method was tested using breast cancer miRNA and mRNA expression bulk data. In both cases, the results showed that the method outperformed the existing methods. In addition, "GO biological processes" and "KEGG pathways" analyses were performed on the miRNA-mRNA causal relationships inferred by the method. Such analyses revealed a significant number of miRNAs/mRNAs associated with cancer and EMT processes, indicating the utility of the method.

Although computational methods may never completely replace intervention experiments, the method developed in this thesis proved to be an effective tool for understanding regulatory relationships that characterise biological processes. The identified miRNA targets can be used to design wet lab experiments that could unveil new aspects of the dynamics of processes such as cancer progression and EMT. This method can be applied to different datasets to uncover miRNA-mRNA causal regulatory relationships that characterise other biological processes.

The second problem addressed in this thesis involves the *discovery of genes driving biological processes that cause cancer progression*. The traditional paradigm defines cancer-causing genes as those harbouring mutations at a consistently high rate [14, 15]. These mutations are commonly known as "driver mutations" and the genes undergoing them are called "cancer drivers".

One consequence of limiting the definition of cancer drivers to that of the traditional paradigm for years is that our knowledge about driver genes became strongly biased towards mutational events. Nevertheless, recent evidence has shown that driver events are not limited to driver mutations, as they can also

happen at levels other than the genome, such as the epigenome (modifications or marks "onto" the genome) [16, 17]. In this thesis, inference models were found to be able to expand our knowledge about driver genes if such inferences are performed in view of the altered biological processes in cancer.

Mutational approaches for discovering cancer drivers have proven to be useful. However, such approaches are only able to consider a subset of the genes that cause cancer. If a gene is a cancer driver, such a gene (regardless of its mutational status) is dysregulating one or more core biological processes. Therefore, analysing the effects (specifically the causal effects) of the dysregulation that occurs during cancer progression provides valuable information about the genes that drive cancer.

To solve this second problem, a causal method for the discovery of cancer drivers (in the sense that genes drive one or more core processes during cancer progression) is proposed. The method overcomes three of the main limitations of current methods used for discovering cancer drivers. First, existing methods primarily use static data, but because cancer drivers alter biological processes, temporal data is required to characterise gene behaviour exhibited during cancer progression. Second, gene mutation data traditionally used for discovering cancer drivers lack information on other essential aspects of cancer evolution. Thus, incorporating other aspects, such as the temporal dimension of the disease, is required to compensate for this limitation. Third, because cancer drivers cause cancer to develop, an approach that can be used to test causality during cancer progression is also required.

Based on the key idea that cancer driver genes cause significant aberrations in core biological processes throughout cancer progression, a method that aims to find cancer drivers by testing the *causal impact* of putative drivers on the biological process is designed in this thesis.

This method was applied to the GSE75688 single-cell dataset, which contains the gene expression of 515 single cells from 11 patients with breast cancer. The method's ability to discover novel cancer drivers and its inference power were assessed by comparing the findings to those of six comprehensive and well-recognized studies. The selected studies provided comprehensive catalogues of highly confident

cancer drivers. Additionally, as most currently available methods for discovering cancer drivers use bulk data, a benchmark with bulk data was also established. The TCGA-BRCA data were used to compare our method with 7 popular methods (which are applicable to bulk data only) for cancer driver inference.

The results were validated by using the Cancer Gene Census (CGC) as a conservative approximation of the ground truth. The results showed that our method outperformed the benchmark methods, as it discovered more CGC genes with higher precision than did the other methods. The experiments suggest that the method can identify genes that drive different altered processes, expanding the range of drivers that can be discovered from data. The results also suggest that the temporal information not only improves the performance of cancer driver inferences, but also allows the genes driving specific processes throughout cancer progression to be identified.

The last problem addressed in this thesis concerns the fact that cancer progression is not a consequence of a single driver gene, but *genes acting together (collaborating and interacting)* in a dynamical system. This thesis's final method addresses two major points. First, causal relationships influencing cancer progression involve sets of driver genes rather than single drivers, but current methods are designed to identify single drivers. In reality, alterations of the biological processes that lead to cancer are caused by the joint dysregulation of several driver genes. Second, current methods neglect the dynamics of gene interactions, and thus models of temporal gene behaviour exhibited during cancer progression remain incomplete.

To overcome these limitations, a method that integrates the temporal dimension of the data throughout the cancer progression and provides a way to test for the causality of candidate gene interactions is developed in this thesis. Specifically, a family of *kinetic models* that explore the causal structure among driver genes in the context of *ordinary differential equations* is proposed. In this approach, the dynamics of gene interactions that occur during cancer progression are explored to shed light on the dynamical behaviour of driver genes throughout cancer progression. These models can be utilised to represent the interactions and collaborations that occur between driver genes throughout cancer progression as a cooperative network. This representation of the joint behaviour of the set of drivers that induce cancer progression (which is not limited to regulatory relationships) provides insights

into the mechanisms that can cause carcinogenesis and the early development of cancer.

This method was applied to the TCGA-BRCA bulk sequencing dataset and the GSE75688 single-cell RNA sequencing data from the NCBI GEO database. The results showed that the method systematically identified *bona fide* driver genes while detecting sets of drivers significantly linked to cancer progression. The experiments conducted suggest that the method can efficiently discover mutated and non-mutated cancer drivers, indicating that it is suitable for datasets in which mutational information is limited. Furthermore, the results suggest that the method is useful for detecting clusters of genes that drive alterations in strongly related cancer pathways.

1.4 Thesis structure

This thesis is structured as follows. Chapter 2 provides a theoretical background on the key concepts that are addressed in this thesis. In Chapter 3, current research methods used in the field are explained. Next, a series of computational approaches developed to overcome the abovementioned limitations are detailed in Chapters 4, 5, and 6. Finally, the conclusions and future directions of the research are discussed in Chapter 7.

Chapter 2

Background

This chapter discusses some the foundational elements that support our current understanding of cancer at the genetic level and their relationship with biological processes in cancer. Section 2.1 explores gene expression and its pivotal role in biological processes, as well as the intricate roles of non-coding RNAs in regulating these processes and their subsequent dysregulation in cancer. Section 2.2 introduces the concept of dynamical systems and stochastic processes, essential for a comprehensive understanding of the complex nature of cancer progression. Section 2.3 shifts focus to the methodologies of causal inference from time-dependent data, crucial for unravelling critical aspects of cancer development. Finally, Section 2.4 presents an overview of cancer data and data transformation techniques, with a special emphasis on data specific to cancer research and the innovative concept of pseudotime data.

2.1 Gene expression and biological processes

Gene expression is the fundamental process that allows all known forms of life, including eukaryotes (such as multicellular organisms), prokaryotes (such as bacteria and archaea), and potentially viruses, to create the fundamental biomolecules necessary for life [18]. Gene expression is the mechanism by which genetic information contained in a gene is utilised to produce such biomolecules, which are commonly referred to as functional gene products [18]. Typically, the process for obtaining functional gene products starts with the *transcription* of genetic information into RNA molecules. In genes that do not code for proteins, such as ribosomal RNA (rRNA), transfer RNA (tRNA), or small nuclear RNA (snRNA) genes, the products are functional RNA molecules. The biomolecules resulting from the *transcription* of protein-coding genes (known as coding RNAs) can be further *translated* into proteins [19, 20].

The basic mechanism that underlies gene expression is explained by the *central dogma of molecular biology* (see Figure 2.1). It stipulates that DNA, the primary repository of genetic information, undergoes *transcription* to encode the genetic information as a complementary RNA molecule. Subsequently, this RNA intermediate serves as a template for protein synthesis in the process of *translation*. Translation pathways operate in a unidirectional manner, and thus prohibit the flow of information from proteins back to RNA or DNA.

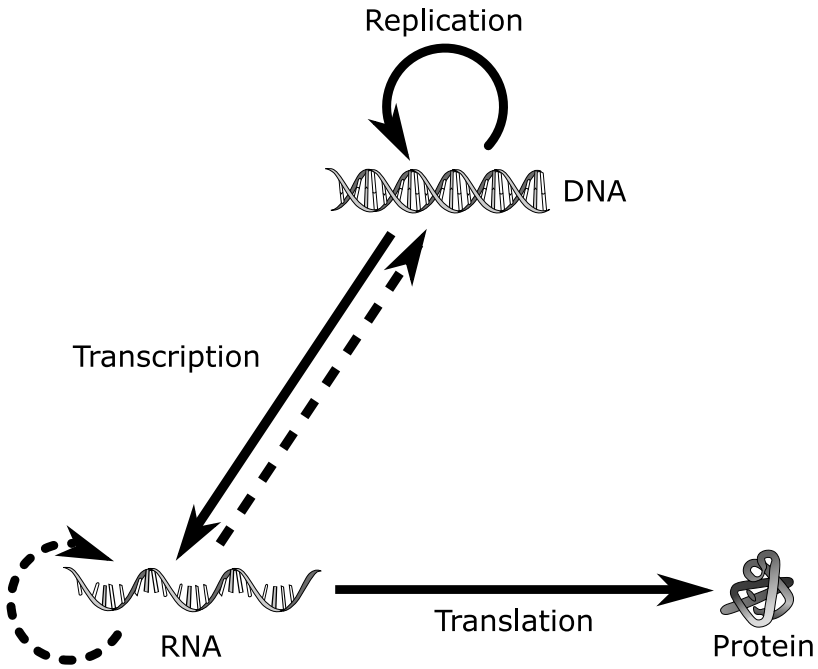


Figure 2.1: **Central dogma of molecular biology.** Graphical representation of the flow of genetic information within cells. Replication: DNA is copied (DNA replication). Transcription: DNA information is transferred into mRNA. Translation: proteins are synthesized using the information encoded in the mRNA. Special transfers are represented by dashed lines. RNA replication: RNA is copied. Reverse transcription: DNA is synthesised using an RNA template. *Author: Philippe Hupé under Licence CC BY-SA 3.0 via Wikimedia Commons.*

Specific gene expression patterns of thousands of genes determine the different phenotypes and functions of cells. These patterns change over time and reflect the balance between the syntheses of gene products and the degradative biochemical pathways within cells. The entire process is highly regulated, and balance is carefully controlled by certain molecules (i.e., regulators) that interact with genes to promote or repress the synthesis of gene products.

The most common point at which gene expression is controlled is at the beginning of the transcription process. At this point, certain RNA molecules such as microRNAs (miRNAs) and certain proteins called transcription factors (TFs) promote or suppress the transcription of genes by binding specific DNA sequences related to such genes. In humans, the number of currently identified TFs exceeds 1600, and they act together to control specific pathways such as immune responses [21].

Gene regulation has far-reaching implications in *biological processes* such as cell dif-

ferentiation, metabolism, and cellular responses to environmental changes. A *biological process* is a series of events that occur in an organism to accomplish a genetically programmed task. Specific sequences of biochemical changes within cells shape such processes. Sequences that lead to the development of biological processes are the result of interactions between basic biological units, such as genes and proteins. Therefore, representing a biological process as a network of interactions between these entities is a common and widely accepted practice. These interactions are causal, which means that changes in one node (i.e., the cause) trigger changes in other nodes after a certain period of time (i.e., the effects).

2.1.1 Non-coding RNA roles in the regulation of biological processes

Historically, the focus of cancer biology research revolved around the functions of protein-coding genes. However, contemporary studies have unveiled that non-coding RNAs (ncRNAs), which are RNA molecules that lack the ability to encode proteins, play significant regulatory roles in cellular dynamics. Specifically, ncRNAs can exhibit cancer-promoting and cancer-inhibiting behaviours. Of the more than 50000 known ncRNAs, the majority have been identified only in the past decade, and many remain largely unstudied. Given these findings, a substantial number of cancer-based clinical trials that leverage ncRNAs as either emergent biomarkers or therapeutic agents are now underway.

Non-coding regions comprise approximately 98-99% of the human genome. Because of their inability to encode proteins, they were initially thought to be merely by-products of transcription [22, 23]. Recent global efforts such as the ENCODE project have shown that vast sections of the non-coding genome are transcribed into numerous RNA molecules [24, 25]. Non-coding RNA transcriptions not only influence foundational biological activities, such as growth and organ functionality, but also play a significant role in various human diseases, especially cancer [26].

In general, ncRNAs are classified according to their size. The ncRNAs that are significant for cancer research include miRNAs, transfer RNA-derived small RNA (tsRNA), and piwi-interacting RNA (piRNA), which tend to be smaller than 31 nucleotides. In contrast, long non-coding RNAs (lncRNAs) are longer untranslated RNAs, with a length surpassing 200 nucleotides (nt). This group of longer ncRNAs

also consists of subclasses such as pseudogenes and circular RNAs (CircRNAs). In cancer research, one of the most well-studied forms of ncRNAs is miRNAs, which are biomolecules that are typically 18-25 nt long and are involved in the regulation of gene expression via RNA interference. These miRNAs can degrade or inhibit the translation function of a target mRNA by, for example, binding to the 3' untranslated region (3'UTR) of such mRNA.

A single miRNA can often modulate multiple targets. Conversely, many miRNAs can target a single mRNA. Furthermore, miRNAs consistently exhibit alterations across all cancer types [27]. These alterations have been pinpointed as influential factors in the modulation of the molecular and cellular dynamics associated with cancer [28]. The role of miRNAs in cancer is yet to be fully comprehended, but typically they seem to act in a dual capacity: as either tumour suppressors or as oncogenes (termed oncomiRs) that promote cancer progression or metastasis [26]. Acting as primary regulators, miRNAs govern the expression of a large number of coding and non-coding genes, including predominant oncogenes such as RAS, MYC, and EGFR, as well as pivotal tumour suppressors such as TP53, PTEN, and BRCA1 [26, 29, 30]. The involvement of miRNAs in the simultaneous modulation of multiple genes indicates that they are important regulators of cellular processes such as development, differentiation, and the stress response.

In addition, lncRNAs have also been discovered to be involved in gene regulation. Unlike miRNAs, lncRNAs do not directly interfere with RNA. Instead, lncRNAs modulate gene expression by interacting with various cellular components such as DNA, RNA, and proteins. They can work either locally near their transcription sites (in cis) or at distant cellular or genomic locations (in trans) [31].

Certain studies suggest that lncRNAs play roles at multiple levels of gene regulation, including epigenetic, transcriptional, post-transcriptional, translational, and even post-translational modification processes. Key functions at the chromatin level include dosage balancing, genomic imprinting, and chromatin modification and restructuring [32]. Furthermore, lncRNAs contribute to the regulation of alternative gene splicing, cellular differentiation, and the cell cycle. Moreover, lncRNAs can act as molecular scaffolds that organise different regulatory molecules and proteins to facilitate vital cellular interactions.

In recent years, it has been found that lncRNAs, such as Hotair, execute their functions by attaching to the proteins responsible for transcription regulation, including those involved in histone modification and chromatin structure alteration. They can also influence mRNAs via direct binding and interact with miRNAs. Through this interaction, which is frequently termed the "sponge effect", an lncRNA can control an miRNA's ability to inhibit gene expression [26].

Overall, non-coding genes are intricately involved in a variety of cellular functions, ranging from directing the specialised roles that cells assume during differentiation to safeguarding the structural integrity of DNA. Although ncRNAs (which are products of non-coding genes) do not code for proteins, they significantly influence the structural dynamics of the genome by either amplifying or suppressing the expression of specific target genes.

2.1.2 Dysregulation in cancer

Dysregulation, defined as an abnormal regulation of gene expression, leads to alterations in biological processes that may upset the balance of critical cellular functions, such as cellular growth. The dysregulation of biological processes is responsible for a wide range of genetic diseases, including Huntington's disease, cystic fibrosis, hemophilia, and cancer [33]. In particular, abnormal gene regulation confers on cancer cells the *hallmarks of cancer*, which are crucial for tumorigenesis and cancer progression.

Dysregulation due to genetic aberrations (e.g. gene mutations) is a critical factor in the development and progression of cancer, as it can alter the normal functioning of genes and lead to the formation of cancer cells [34]. Such aberrations can induce dysregulation in cancer cells by promoting the activation of oncogenes or stifling the expression of tumour suppressor genes. Oncogenes are genes that normally regulate cell growth and division, but when they harbour genetic aberrations, the overexpression of such genes may promote uncontrolled cell division and tumour formation [34]. For example, mutations in the oncogene KRAS that encode a small GTPase protein and regulate cell signalling pathways involved in cell growth, differentiation, and survival have been observed in approximately 30% of all human cancers, including pancreatic, colon, and lung cancers [35].

Underexpression of tumour suppressor genes also contributes to cancer progression. Tumour suppressor genes normally function to prevent uncontrolled cell division and promote cellular death. Inactivation of these genes can occur via point mutations, deletions, or chromosomal rearrangements, leading to a loss of their tumour suppressor function [36, 34]. One of the most well-known tumour suppressor genes with mutations associated with cancer is the TP53 gene. TP53 is a gene that encodes the p53 protein, which regulates the cell cycle and triggers apoptosis in response to DNA damage. Inactivating mutations in TP53 have been observed in a variety of human cancers, including lung, ovarian, and colorectal cancers [37].

The dysregulation of gene expression can also contribute to the development and progression of cancer through the alteration of DNA repair pathways. DNA repair genes are responsible for correcting errors that occur during DNA replication, and they play a critical role in maintaining the integrity of the genome. Therefore, defects in these pathways can result in genomic instability and an accumulation of aberrations [38, 39]. For example, mutations in the DNA repair genes BRCA1 and BRCA2 lead to an increased risk of breast cancer [40].

The dysregulation of miRNAs is another well-established genetic mechanism involved in cancer initiation and progression. These small non-coding RNA molecules regulate gene expression by binding to mRNA and preventing it from being translated into proteins [41]. Studies have shown that several miRNAs are differentially expressed in cancer cells relative to normal cells [42]. For instance, some miRNAs may be overexpressed in cancer cells, leading to the repression of tumour suppressor genes and the uncontrolled cell growth and division that is characteristic of cancer [41]. Conversely, some miRNAs may be underexpressed, resulting in the activation of oncogenes and the facilitation of tumour progression [43].

Additionally, changes in miRNA expression can also affect the functioning of immune cells, leading to a more permissive tumour microenvironment [41]. The immune system plays a crucial role in the recognition and elimination of cancer cells. Alterations in miRNA expression can lead to the suppression of the immune response, thereby promoting tumour growth and progression [44]. As therapeutic targets, miRNAs have received considerable attention in recent years because of their major role in cancer. By restoring normal miRNA expression or blocking

the function of disease-promoting miRNAs, new drugs have the potential to treat cancer at the molecular level. For example, a recent study showed that the restoration of the normal levels of mir-34a in cancer cells led to the inhibition of tumour growth and increased apoptosis [45].

Although gene dysregulation in cancer is traditionally linked to gene mutations, aberrations other than mutations can grant the same capabilities to cancer cells. For example, epigenetic alterations have a significant impact on gene expression. Epigenetic alterations are chemical changes to DNA or the proteins linked to it that modify the accessibility of the genetic material without changing its underlying nucleotide sequence. For instance, adding methyl groups to DNA cytosine residues results in DNA methylation, a well-known type of epigenetic alteration that limits access to the underlying genes by condensing the chromatin structure.

A decreased level of DNA methylation in tumours relative to normal tissue has been frequently observed in human cancer [46]. This loss of methylation can be attributed primarily to the hypomethylation of repetitive DNA sequences and the demethylation of coding regions and introns, which are regions of DNA that control alternative versions of mRNA transcribed from a gene [46]. The consequences of DNA hypomethylation are frequently unpredictable and can vary significantly. For example, while hypomethylation may suppress the advanced stages of tumour development in the intestine, it may simultaneously promote early precancerous lesions in the colon and liver [47].

Regardless of the underlying mechanisms that induce the onset and progression of cancer, gene dysregulation that alters biological processes is a common element in the genetic causes of cancer. This key fact and the fact that biological processes emerge from dynamic interactions between genes are used as starting points in this thesis. By carefully considering the *gene dysregulation that induces alterations in biological processes* during the progression of cancer, a better understanding of the dynamics of the disease can be gained.

2.2 Dynamical systems and stochastic processes

Throughout this thesis, a *system* is defined as the combination of elements (or components) acting together to achieve a specific objective [48]. Elements that shape a

system are single functioning units (e.g. genes). Although this concept of *system* is frequently employed to describe physical phenomena, the same concept applies to abstract phenomena, which are typical in fields such as economics [48].

If the output of a system depends on past inputs, such a system is considered *dynamic*. In contrast, if the output of a system depends exclusively on current inputs, such a system is considered *static* [48]. One important characteristic of dynamic systems is that the output of the system changes with time if the system is not in equilibrium [48, 49]. One implication of this phenomenon is that changes in the constituent elements of the system can be reflected in its future states, which does not occur in static systems.

In the field of dynamic systems analysis and modelling, the focus is often on deterministic systems. In these systems, the future outputs of the system are not considered random variables, and variations in the inputs and outputs of the system are usually modelled as low-intensity noise [49]. Despite the usefulness of deterministic approaches in different areas of knowledge such as engineering [49], the direct application of these approaches at the biological level is impractical for two reasons. First, the number of variables involved in biological dynamic systems exceeds their engineering analogues by several orders of magnitude. Second, biological variables are usually considered random variables because of their probabilistic nature, and hence modelling the variations observed in the system as noise is insufficient.

This thesis focuses on gene entities as system components and their interactions as *dynamic systems*. Specifically, alterations in the components and dynamic interactions of the system that endow cancer cells with the ability to initiate the processes of tumorigenesis and cancer progression are examined. Furthermore, the components of biological dynamic systems are considered random variables and the manifestation of dynamic interactions over time (gene combinations) are considered biological processes. Therefore, these analyses are classified under the category of *stochastic dynamic systems*.

Considering stochastic nature of biological systems, the definition of dynamic systems includes the set of stochastic processes in which a set of random variables evolves over time. Formally, a *stochastic process* (also known as a random process)

can be defined as a collection $\mathbf{X}$ of random variables indexed by a parameter t (often used to represent *time*) that belongs to a set T (i.e., $\mathbf{X}(t), t \in T$) [50]. Strictly speaking, a biological process is continuous, and thus the expression of genes involved in it is a random process that can be expressed by a common probability model with an infinite collection of random variables (because there are an infinite number of points in time in an interval) [51].

To provide an intuitive conception of the infinite number of random variables involved in a random process, consider the gene expression profiles of a population P of n individuals. Theoretically, it is possible to collect gene expression data on the genes related to a biological process for all individuals in the population P at a particular point in time t, which would enable a gene expression distribution to be constructed for each gene. Thus, genes at instant t can be considered random variables, and their numeric measurements at time t can be considered their expression. Because the number of points in time is infinite, the number of random variables associated with each gene is also infinite.

Biological systems (as well as a large number of other systems) do not change abruptly from one instant to the next. This "inertial" characteristic of biological systems allows information on the dynamics of these systems to potentially be recovered from the sequences of observations (known as time series [52]) of that system's variables. In other words, the most important aspects of a biological process can usually be described in terms of a finite collection of sequential data. Regardless of the classification of the dynamic system (deterministic or stochastic), one of the main goals of dynamic system analysis is accurate modelling.

One way to achieve a sufficiently accurate model of a dynamic system is to apply physical and/or mathematical laws to it [48]. However, in many systems (such as biological systems), this analytical strategy may be infeasible because of the complexity of the problem. Another approach for modelling a dynamic system is to infer the dynamic characteristics of the system from its observed behaviour [52]. From this perspective, the problem is transformed into a *reverse engineering problem* in which models are developed based on domain expertise and available data, and they are assessed by using rigorous statistical models or model-checking methods [53]. This reverse engineering strategy is adopted in this thesis.

The research conducted for this thesis revealed that the characterisation of relevant aspects of the dynamics of cancer can be accomplished by applying data-driven approaches to sequential data. In this thesis, solutions that overcome certain limitations of traditional approaches for the modelling and analysis of dynamic systems (which cannot be applied directly) are developed in order to analyse the underlying dynamics of cancer. These solutions are carefully described in Chapters 3, 4, and 5.

2.3 Causal inference from temporal data

There is currently increasing interest in the field of causal structure learning, particularly in methods to infer causal relations from data. This is a response to the fact that information on the causality of different phenomena in a significant number of areas is more useful than information on associations (e.g. correlations). Many methods that aim to identify causal structures have been developed [54, 55, 56, 57, 58, 59]. However, these methods either depend on strict assumptions that are often not upheld when processing heterogeneous data, or they do not produce a single causal graph owing to general identifiability problems [60].

The (probabilistic) causal models shown in Figure 2.2 (adapted from [61]) are two examples of the (no minor) identifiability problem. These models represent causal relationships between random variables, as indicated by the arrows, which are directed from the causes to the effects. The parameters required to fully define these probabilistic causal models are the probabilities of the variables with no causes and the conditional probabilities of the effect(s) given its/their cause(s). Thus, the parameters for the first model (Figure 2.2(a)) are $P(a)$, $P(b|a)$, and $P(c|b)$, while the parameters for the model in Figure 2.2(b) are $P(b)$, $P(a|b)$, and $P(c|b)$. Because $P(a)P(b|a) = P(ab) = P(b)P(a|b)$, the values obtained for the parameters in the first model completely determine the values of the parameters in the second model, and vice versa.

Traditional probabilistic causal modelling (as demonstrated in the previous example) is defined in terms of the probability distribution of the variables in the system, overlooking the temporal precedence between causes and effects. Relying exclusively on the probability distributions of the observed data may not be sufficient for data-based identification of the true causal structure of most systems. The work

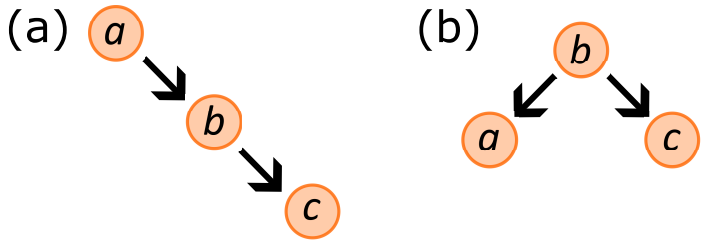


Figure 2.2: **Example of two equivalent causal models**. Equivalence can be observed in terms of the independence information conveyed by the corresponding *Directed Aciclyc Graphs* (DAGs). These two graphs represent the independence statement $I(a, c|b)$, which is read as "a is independent of c given b". Example adapted from [61].

performed for this thesis revealed that including a temporal dimension in the data assists in palliating the identifiability problem of causal structures in data-driven approaches. However, the temporal precedence of the variables describing a system is a necessary but not sufficient condition for demonstrating causality.

Although temporal precedence alone does not determine causation, the difficulty in determining the causes and effects can be significantly reduced by considering time series data. The inclusion of a temporal order in the data increases the confidence of causal discovery because of the temporal precedence between cause and effect. In a causal system that progresses over time, changes in the causes happen before the changes in the affected variables. This concept has been explored in approaches such as *Granger causality* [62] and *structural vector auto-regressive models* [63].

Granger causality explores the idea of causal relationships between time series signals in terms of the utility of one or more time series for predicting another time series of interest. A *Granger cause* can be briefly formulated as follows: if a signal y causally depends on a signal x, then the future state of y can be more accurately predicted if the information on x is taken into account [64]. Accordingly, x is the *Granger cause* of y if using the information on x improves the predictability of y. Despite its usefulness, one of the main criticisms of *Granger causality* is that it fails to capture the "structural causality" (i.e. a *Granger cause* is not necessarily a cause) [65].

Another popular suite of modelling methods that consider temporal precedence are called *structural vector autoregressive* (SVAR) approaches. SVAR approaches as-

sume that a system can be modelled by regressing each model variable with its own lagged values, the lagged values of the other model variables (*vector autoregressive* or VAR), and the instantaneous relations between the model variables [66]. An SVAR model can be thought as a VAR model that also considers the contemporaneous causal structure [65]. One of the main drawbacks of these approaches is that they are designed to detect the causal effects of exogenous (external) variables (usually named *shocks* in the field of economics) rather than the causal relationships between the endogenous (internal) variables of the system [67]. In this thesis, finding the causal relationships between internal variables (e.g. genes) that alter biological processes in cancer are of interest.

Probabilistic causal methods and temporal precedence modelling methods have complementary advantages. Therefore, a number of methods utilise aspects of one or both approaches to make causal inferences from time series data. In the following subsections, critical concepts utilised in the solutions derived in this thesis are discussed.

2.4 Cancer data and data transformation

2.4.1 Data for cancer research

The international cancer research community continuously generates extensive quantities of molecular and phenotypical data, with the goal of understanding the intricate mechanisms that drive carcinogenesis and cancer progression [68]. Examples of international projects dedicated to the collection of tremendous amounts of cancer data for scientific purposes include The Cancer Genome Atlas (TCGA) [69], which is comprised of more than 2.5 PB of raw data, and the Gene Expression Omnibus (GEO) database [70], which has more than 1.1 million biological samples under the "cancer" category. These datasets contain a substantial amount of information that can provide innovative perspectives on fundamental inquiries. Because of this prodigious volume of cancer data, processing them requires "large data", or even "big data" techniques [68].

The most common types of data in cancer research can be divided into the categories of molecular omics, perturbation phenotypical, molecular interaction, imaging, and textual data [68]. The most abundant type is molecular omics data, which

describe the abundance and status of biomolecules in cellular systems and tissues. One of the primary reasons why omics data comprise a significant portion of the cancer data used in research is that they can be obtained directly from patients or pre-clinical samples.

The methods developed in this thesis focus on one particular category of omics data known as transcriptomics (specifically, gene expression levels). This category was chosen because the dysregulation of biological processes are reflected in abnormal gene expression patterns. Table 2.1 displays a summary of the most common types of molecular omics data.

Table 2.1: **Common types of molecular omics data in cancer research**. Most technologies used for collecting molecular omics data process *bulk data* by default. Adapted from [68]

Data category	Examples of data	Typical technology
Genomics	DNA mutations	Whole-exome/ whole-genome sequencing
Epigenomics	Chromatin accessibility	ATAC-seq or DNase-seq
	Histone modification	ChIP-seq
	DNA methylation	Bisulfite sequencing and BeadChip
Transcriptomics	Gene expression level, transcript isoforms	Microarrays, RNA-seq
	Gene expression data with spatial location information	Spatial transcriptomic techniques
Proteomics	Protein expression and phosphorylation	Mass spectrometry, protein array
	Cell surface proteins	CITE-Seq
	Physical, chemical, and fluorescent characteristics of cells (or particles)	Flow cytometry
	Antibody binding to cellular antigens	Mass cytometry
Metabolomics	Metabolites from patient samples	NMR spectroscopy, mass spectrometry

In general, most current technologies work on tissue samples known as bulk samples, which usually contain a large number of cells. The collective gene expression measurements obtained from a population of cells are commonly known as *bulk data*. Data from bulk samples have been widely used in genomics research to study gene expression patterns and to provide insights into cancer development

[71, 72], tumour evolution and metastasis [15], and cancer drivers [73, 74, 75]. In bulk RNA sequencing (RNA-seq), thousands of cells are lysed together, and their RNA is sequenced as a pooled sample.

RNA-seq provides an average gene expression profile for the entire population. Bulk data analyses have proved to be valuable tools for studying large cohorts of patients. However, conventional bulk data analysis lacks information about individual cells. Therefore, its ability to characterise the heterogeneity of cancer is limited. In addition, bulk data primarily produce summary statistics of the cell populations in tissue samples. Consequently, bulk data mask a large amount of information on cancer development and progression.

Owing to the development of sequencing technologies, a wide range of molecular data can also be generated at single-cell resolution [76, 77]. These single-cell data, as the name suggests, allow the analysis of gene expression at the resolution of individual cells. Furthermore, single-cell RNA sequencing (scRNA-seq) technologies can produce high-throughput profiles of transcriptomes, which can provide insights into cellular heterogeneity and dynamics [78, 79]. Single-cell data can probe tumour heterogeneity and evolution at a very high resolution, allowing for the identification of cell-to-cell variability and the stochastic gene expression landscape within complex biological systems.

Although single-cell data are promising because of their high level of detail, they have several technical and scientific limitations that still await resolution. Producing cell maps with a very high level of detail or precision requires significant methodological effort. Such a level of detail is usually limited by the degree of flexibility and constraints imposed by the experimental setup during the data collection process [80]. Other issues such as noise control, missing data, and high levels of sparsity remain open challenges in single-cell data research [81].

2.4.2 Pseudotime data

One ambitious goal in cancer research is to describe the dynamical behaviour of the cellular mechanisms that lead to cancer. This goal requires unveiling the dynamics of gene regulation [82], which is one of the central aspects of this thesis. To understand the dynamics of gene regulation in cancer, data with a temporal dimension (e.g. a time series) is required. Because obtaining molecular omics data

is destructive to cells, RNA-seq/scRNA-seq data are static representations of the cellular states of tissue samples. Consequently, real time series data are not easy to obtain in practice. However, in this thesis, a well-recognised data transformation called *pseudotime* is adopted as the temporal dimension of the data.

The dynamic progression of cellular states can be indirectly estimated from population snapshots using methods that model a curve or a tree based on the continuous distribution of cells in a high-dimensional state space [82]. Estimation of such progression is known as *trajectory inference* (TI), or *pseudotime* analysis. *Pseudotime* inference methods are computational approaches that assign a pseudotemporal order to samples (usually cells) based on similarities in their expression patterns [83]. Although *pseudotime* analysis is primarily employed in the context of single-cell data, recent studies have shown that *pseudotime* is capable of extracting a portion of the underlying dynamics of cancer progression, even from bulk data [84, 85, 78, 86, 87].

Pseudotime analysis aims to assign a series of one-dimensional quantities, known as pseudotimes, to a collection of high-dimensional molecular data obtained from a cross-sectional cohort of individuals (Figure 2.3). By this means, a *pseudotime* approach quantifies the relative progression of samples along a biological trajectory [84]. In this way, *pseudotime* analysis enables the reconstruction of dynamic processes, such as cell fate decisions, lineage commitment, and cellular transitions.

Pseudotemporal order is intrinsically temporal and it reflects the progression of the biological trajectory, as encoded by the inferred *pseudotime* [88]. To estimate pseudotime scores, a plethora of methods has been developed over the past few years [84, 88, 89, 85, 90, 91]. These algorithms utilise different strategies, such as diffusion maps, principal component analysis (PCA), minimum spanning trees, and more, to capture the dynamics contained in omics data [83]. Regardless of the underlying strategy, *pseudotime* methods map each sample onto a one-dimensional pseudotemporal progression (or set of latent states). This score can be used to order the gene expression by following the *pseudotime* in ascending order. Additionally, if the cohort of samples belong to different sub-populations, the *pseudotime* can reveal different temporal patterns for the same gene, which is critical for understanding how the dynamics of genes change as cancer develops.

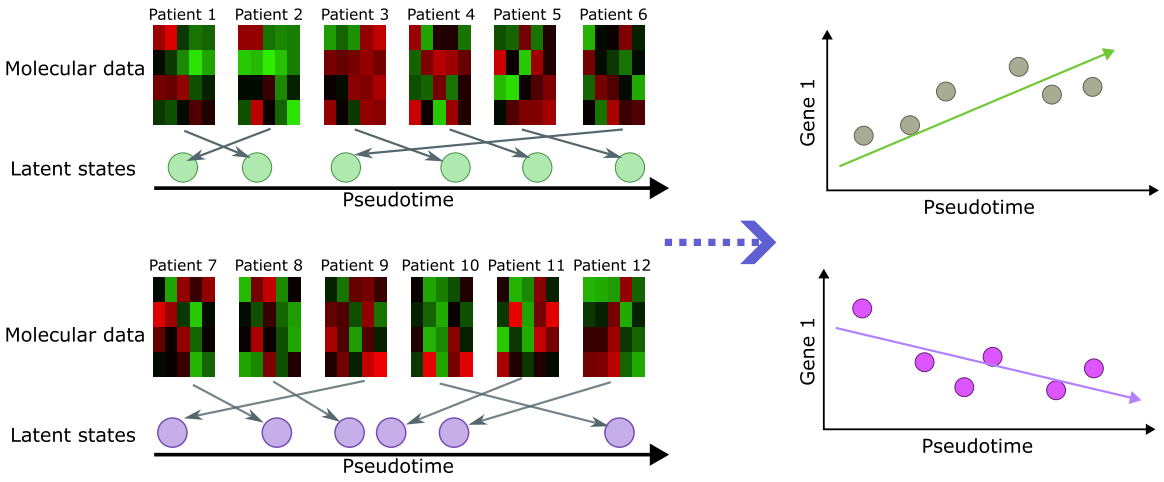


Figure 2.3: **Illustration of pseudotemporal analysis.** High-dimensional omics data (usually gene expression) are mapped onto a one-dimensional feature (*pseudotime*). *Pseudotime* represents the latent state of samples along a biological transition. For cohorts with subpopulations (e.g. different phenotypes), *pseudotime* analysis can uncover different temporal gene expression behaviours for the same gene. Our research has shown that *pseudotime* is useful for capturing the underlying dynamics of biological processes involved in cancer from both bulk and single-cell data. Adapted from [84]

.

The research performed for this thesis revealed that incorporating *pseudotime* into causal analysis provides valuable insights into cellular dynamics and regulatory processes because it enables the identification of pathways associated with developmental processes and biological transitions. Therefore, in this thesis, this capability is utilised and combined with elements of causal inference to detect the regulatory relationships that characterise biological processes involved in cancer (Chapter 4).

Furthermore, by ordering samples along a trajectory, *pseudotime* analysis facilitates the identification of critical points in time and transitional states that occur during cellular processes. The research conducted for this thesis showed that the temporal aspect of the *pseudotime* score can be used for detecting cancer drivers. In the approach proposed in this thesis (Chapter 5), cancer driver genes are identified by performing a *counterfactual analysis* to detect the *causal impact* of a driver after a critical point in which the transition from a normal to a carcinogenic state for a core biological process is irreversible.

Finally, the *pseudotime* approach and a dynamic casual inference approach are combined to detect the set of genes that interact and collaborate to cause cancer to de-

velop. By performing a data transformation based on *pseudotime*, the rate of change of such interactions and collaborations can be modelled as a *dynamic system*. The details of this approach can be found in Chapter 6.

Chapter 3

Literature review

This chapter presents a literature review on data-driven methods employed in the inference of gene regulatory networks and the discovery of cancer drivers, essential for unravelling the complex biological mechanisms underlying cancer dynamics.

Section 3.1 is dedicated to gene regulatory networks inference methods, starting with co-expression methods (Section 3.1.1) that investigate gene expression correlations. It then progresses to Machine Learning approaches (Section 3.1.2), highlighting their role in interpreting genetic interplays. Section 3.1.3 discusses Differential Equation Based Methods for modelling gene regulatory dynamics, followed by a thorough examination of causality based Methods in Section 3.1.4, which includes Granger causality, Bayesian-based methods, and other causal approaches, underlining their significance in establishing gene regulation causality.

Section 3.2 explores cancer drivers discovery methods, beginning with mutation based methods (Section 3.2.1) that delve into the genetic mutations' role in cancer, covering frequency, hotspot, and functional impact based approaches. This section is extended with network based methods (Section 3.2.2) for integrating complex biological networks in identifying key cancer drivers, and further with causality based identification of cancer drivers (Section 3.2.3), emphasising the application of causal inference in pinpointing cancer drivers.

Finally, Section 3.3, gene group analysis of cancer drivers, examines the collective behaviours and characteristics of gene groups implicated in cancer, providing insights into the collective impact and interactions among cancer-associated genes. This chapter aims to offer an overview of the current methods in genetic regulatory network inference and cancer driver discovery, providing further context to the research developed in this thesis.

3.1 Methods for inferring gene regulatory networks

The temporal patterns of gene products shape the different biological processes that ultimately govern biological systems and functions in organisms [92, 93]. Gene regulatory networks (GRNs) describe a collection of molecular interactions that jointly control the abundance patterns of gene products [93]. GRN topology is highly complex because a vast number of interdependent molecular interactions

are involved. The regulatory complexity itself represents a significant challenge in accurately modelling biological systems [92]. However, because of the relevance of GRNs, identifying the mechanisms responsible for gene expression levels and their underlying structure (i.e. GRN inference) has been an exhaustive area of research for several decades [94].

The process of GRN inference starts with the input data, which are typically gene expression data obtained from either experimental procedures or databases. In a preprocessing stage, computational inference algorithms use the input data to reverse engineer the network structure. Next, the inferred model is usually further optimised and compared with curated biological knowledge to obtain valuable new information [95]. Owing to advances in bulk and single-cell profiling technologies, it is now possible to obtain further insights into the molecular and cellular actors involved in biological processes. However, datasets obtained by using current technologies are high-dimensional, and thus their analysis requires a series of specialised computational methodologies [92, 94, 96].

The most common practice in GRN inference is to assume that the network structure can be represented as a directed graph $G(X, F)$. In this representation, X is a variable (node) linked or related to other variables through a function F that usually weights such connections. In a directed graph, connections between nodes (called edges) are usually represented as arrows pointing from a parent node to a child node. In the context of GRNs, nodes are usually genes or relevant biomolecules such as mRNAs, miRNAs, and proteins.

A significant number of computational methods have been developed to characterise regulatory interactions that can be derived from genomics data [97, 98, 99, 100, 101, 102, 103]. Depending on the way in which the methods identify regulatory relationships, these methods can be categorised as co-expression methods, machine learning methods, differential equation-based methods, and causal-based methods.

Even though a number of current GRN inference methods use sequential [104, 102]data, they do not formally describe the dynamics of causal relationships between genes. In certain cases, the relationships between the network nodes are quantified by measures of association that do not provide information about which

genes are influencing and which genes are being influenced. Various methods use time as a proxy to determine the direction of the influence. Although changes in the cause occur before changes in the effects, simple temporal precedence is insufficient to determine a causal link or to determine the dynamics of the relationship itself.

Additionally, current methods cannot fully characterise the regulatory relationships that occur *during* cancer progression. This is because GRN inference performed for cancer research is currently conducted using methods that are not specifically designed for cancer. Therefore, they do not account for the alterations in central biological processes that characterise the disease, and thus they likely overlook the dynamics of regulatory relationships that cause such alterations.

This thesis was motivated by the aspiration to overcome these limitations because analysing the gene interaction dynamics that occur throughout the progression of cancer is key to its complete understanding. In this sense, GRN inference must be accomplished by accounting for the dynamics that govern the alterations of the biological processes that occur during the disease.

In the following subsections, the main categories of GRN inference methods are discussed to convey the current state of the subject. The most influential methods are presented, and methods involving a temporal dimension in the data are emphasised.

3.1.1 Co-expression methods

Co-expression GRN inference methods, frequently named *information theory-based* methods, aim to establish regulatory relationships via measurements such as *Pearson/Spearman/Kendall* coefficients, *Euclidean* distances, or *mutual information*. Owing to the simplicity and low computational cost of these methods, they are especially useful for discovering large GRNs in a small number of data samples [95]. However, the similarity measures they use are scalars, and thus the relationships inferred through these methods are undirected. Because regulatory relationships are causal, a full characterisation of a GRN requires the identification of the cause-effect pair involved in the regulatory relationships.

Despite this limitation, several methods based on co-expression are being used as references for new developments. One of the classical approaches in this category is the REVerse Engineering ALgorithm (**REVEAL**) [97]. This method analyses the *mutual information* between pairs of inputs (perturbations or internal changes) and outputs (state transitions or temporal responses of gene expression) in networks. The original research only considered Boolean networks, which simplified the networks inferred from the gene expression data to 0 (unexpressed) or 1 (expressed). The idea behind REVEAL is simple, and it has provided a basis for other methods that are scalable to more complex Boolean models.

In Boolean methods such as REVEAL, a network state S and the outputs of all nodes produced in the same instant are defined as a Boolean function of their respective parents. Depending on the way in which the set of functions is declared, these models are able to present a simplified version of the GRN dynamics [95, 105]. **MIBNI** [101] and **GABNI** [106] are two examples of Boolean network inference methods that use *mutual information* for the inference problem. Both methods utilise Boolean time series discretised from gene expression data to calculate a Boolean function that estimates the binary state of a gene as a function of its immediate past state. In addition, GABNI incorporates a genetic algorithm that searches for an optimal regulatory function.

Boolean modelling partially decreases the computational complexity necessary for inferring large networks. Nevertheless, this simplification of gene expression data into binary values relies on the assumption that gene expression can be described as fully activated or fully inhibited, which is rarely valid. Furthermore, the inferred regulatory relationships are prone to a large number of spurious discoveries because Boolean approaches do not test for explicit causality.

The Algorithm for the Reconstruction of Accurate Cellular Networks **ARACNE** [98] is considered the pseudo gold standard by several contemporary researchers. ARACNE is based on the assumption that biological networks can be approximately modelled as two-way networks. Thus, all of the statistical dependencies can be obtained from pairwise marginal distributions. By using the *mutual information* framework, ARACNE determines the putative gene pairs that interact in the inferred GRN. The method uses the *data processing inequality* (DPI) to remove indirect candidate interactions (false positives). This additional step is performed

because *mutual information* is prone to treat both direct and indirect regulatory relationships in the same way.

Various ARACNE-based algorithms have been developed over the past decade, and a number of them are intended to be used with time series data. One example of such an algorithm is **TimeDelay-ARACNE** [99], which uses the pairwise information between time series that describe gene expression. The method finds the *mutual information* between one gene and the delayed version of another gene. It also uses the delay as a proxy for the direction of the relationship, and thus it does not necessarily represent real causality.

The Fuzzy NEtworks for modelling of gene association method (**FyNE**) is a method based on fuzzy logic and *Kendall* correlation, and it uses previous knowledge to infer the structure of a GRN [100]. The correlation rank is calculated for each gene expression pair in the input dataset to obtain the rank correlation vector **g**. The GO database is employed as the previous knowledge to calculate the strength of the dissimilarity $GFD(g, g')$ between two genes denoted as g and g'.

In FyNE, the fuzzy algorithm receives both the correlation rank and strength of the dissimilarity. The method returns an undirected graph, the elements of which are represented by the triplet $(g, g', f(g, g'))$, where $f(g, g')$ stands for the score of the relationship as calculated by the fuzzy aggregator. The final output is produced after the inferred GRN is trimmed according to a minimum valid link value (α) for $f(g, g')$.

The Lag-based Expression Association for Pseudotime-series **LEAP** [107] uses the *Pearson* correlation coefficient to determine the existence of a regulatory relationship. The method calculates the correlation of normalised mapped-read counts over fixed temporal windows with different lags. The maximum values found by the method are then used to construct the network. In this method, the correlation between two genes is not symmetric because the time windows are included. LEAP uses this asymmetry to orientate the edges in the inferred network.

3.1.2 Machine learning methods

Machine learning methods transform the problem of GRN inference into prediction or classification [108]. Various machine learning methods aim to predict the

response of a gene as a function of other genes (as a regression problem). Other methods assume the identification of a regulatory relationship is a classification problem. For example, they define a class label as regulation (TRUE) or not regulation (FALSE).

These methods are able to model functional relationships contained in the data, and they can handle both static and time series profiles of gene expression. Because of their adaptability, these models are useful for inferring complex behaviour, including that of non-linear systems. Nevertheless, owing to their computational complexity, they have been applied only to very small systems. In practice, describing the behaviour of GRNs in cancer to a sufficient degree of accuracy is challenging because of the fact that real regulatory networks characterising the disease are usually large.

One of the most prominent machine learning methods is GEne Network Inference with Ensemble of trees (**GENIE3**) [109], which uses ensemble tree-based methods to assign one decision tree to each gene to be explored. Each tree infers the regulatory relationships by transforming the problem into a feature selection problem. The method aims to predict the likelihood of a regulatory link between genes by evaluating the utility of the expression of one gene (the regulator) for predicting the expression of another gene (the regulated). GENIE3 infers directed regulatory networks, but it does not test for causality or model the dynamics of the regulation.

Another example of the successful implementation of machine learning methods for GRNS inference is the **extreme machine learning approach** [110], which builds a neural network model for each possible combination of regulator-regulated genes. The training error obtained after several training epochs is used as the parameter that determines whether or not a gene-gene relationship is feasible. The method was tested with both in silico and real data with a low number of genes (up to 50 genes for the in silico datasets and between 5 and 11 genes for the real data). Despite the fact that its performance is apparently superior to similar inference tools, its application to GRN identification in cancer does not seem feasible owing to its low scalability.

GRNBoost2 is a method based on GENIE3. It executes a regularised *regression per target* GRN inference by using a gradient boosting machine to model regres-

sion trends in static data [111]. The inference process can be parallelised in order to mitigate the scalability issue common in machine learning methods. Each regression model is considered a partial GRN, and the final result is obtained by assembling all the regression models. Similar to GENIE3, GRNBoost2 is designed to infer regulatory relationships between transcription factors (TFs) and regulated genes. GRNBoost2 can be used for finding a number of regulatory interactions in cancer. However, its utility is limited because regulatory relationships in cancer are not solely characterised by TF interactions.

The **scTenifoldKnk** method [112] identifies regulatory relationships by simulating gene knockouts, which are called *virtual KO screens*. The method infers an initial GRN from the data as a baseline for further comparisons. Next, the method uses the adjacency matrix (i.e. the matrix representation of the network) and a combination of principal component regression and tensor decomposition to create versions of the network in which a gene has been "knocked out". A comparison between the networks enables the differentially expressed genes that occur as a result of the gene knockout to be detected. The network is created by assigning a directed link between the virtually knocked out gene and the differentially expressed genes that are triggered by the knockout.

3.1.3 Differential equation-based methods

Models based on ordinary differential equations (ODEs) seek to more accurately approximate the dynamic behaviour of the actors in a GRN. In this approach, a set of differential equations is defined to represent changes in the expression of one gene as a function of the expressions of other genes. This representation is usually expressed as

$$\frac{dx_i}{dt} = f_i(x_1, x_2, ..., x_n, p, u), \tag{3.1}$$

where $(x_1, x_2, ..., x_n)$ represents the expression levels of all genes except x_i, u is an external parameter that represents perturbations in the system and/or noise, and p is the parameter set of the system [113]. Although this approach explicitly addresses the dynamics of the problem by virtue of its conception, its formulation for GRNs is underspecified. For this reason, the system analysis is commonly restricted to discovering a linear approximation around a unperturbed steady state.

However, this method ignores transient phenomena, which form a significant component of the dynamics. This issue is particularly critical in cancer research because the transient states of cancer correspond to its early stages. Understanding the dynamics of the early stages of the disease is crucial for early detection, which can significantly improve the prognosis. In addition, ODE methods are unable to be scaled to process the large amount of data required to recover the dynamics of a biological system.

The reverse engineering schema for large-scale dynamic regulatory network reconstruction **HiDi** [114] is an example of a method that attempts to solve the scalability problem inherent to techniques based on ODEs. The method produces the GRN representation as a directed graph $G = (V, E)$, with nodes (genes) denoted as $V = \{v_1, \ldots, v_n\}$ and directed edges (regulatory interactions) denoted as $E = \{e_1, \ldots, e_m\}$. Regulatory networks are inferred via the ODE

$$\mathbf{Y} = \frac{d\mathbf{x}}{dt} = \mathbf{a_0} + \mathbf{A}\mathbf{x}, \tag{3.2}$$

where $\mathbf{x} \in \mathbb{R}^{n\times 1}$ is the gene expression vector and $\mathbf{a_0} \in \mathbb{R}^{n\times 1}$ contains the interception values. The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ contains the components a_{ii} = the basal synthesis and degradation rate, and a_{ij} = for which $a_{ij} > 0$, $a_{ij} < 0$, and $a_{ij} = 0$ indicate the activation, inhibition, and absence of regulation, respectively, of gene i over gene j.

The scalability of HiDi was sharply improved by using the *Frobenius norm* to convert the parameter identification problem into an optimization problem and by using the *Lagrange multiplier* method to solve the system through a *vectorization operator*. Numerical differentiation was performed by fitting a polynomial to the observed data (details can be found in [114]). To solve the optimisation problem, the method requires a large number of perturbations over the genes of interest. In the original in silico experiments, such perturbation data was created by using the *GeneNetWeaver 3.1.1* software package [115]. However, creating a large number of perturbations in real cancer data is infeasible.

dynGENIE3 [116] is a version of GENIE3 adapted for analysing time series data. The method assumes that the derivative of the expression of gene x_j at time t de-

pends on other genes (denoted as $\mathbf{x}$), and it can be modelled as the ODE

$$\frac{dx_j(t)}{dt} = -\alpha_j x_j(t) + f_j(\mathbf{x}(t)), \forall j, \tag{3.3}$$

where f_j is a function found by using an ensemble of regression trees. In order to learn f_j, the method also requires a static dataset that represents steady-state gene expression. A GRN inferred via dynGENIE3 is created by assembling all functions f_j. The dynGENIE3 method was tested by using DREAM4 in silico data as well as three real gene expression datasets for simple organisms: *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Escherichia coli*.

SCODE [117] is a method designed to identify regulatory relationships between TFs and target genes. The method assumes the expression dynamics of the TFs can be modelled according to the following linear ODE:

$$d\mathbf{x} = \mathbf{A}\mathbf{x}dt. \tag{3.4}$$

The regulatory network is found by optimising the matrix $\mathbf{A}$ via dimensional reduction so that the ODE system can properly represent the observed gene expression data. The dimensional reduction procedure is implemented to reduce the computational complexity of the method. However, this may cause problems with the fitting.

3.1.4 Causality based methods

Causality based methods for GRN inference determine the direction of the gene relationships while simultaneously reducing the number of spurious findings by assessing causality. These methods ascertain the existence of regulatory links in the network by testing different properties of causal structures. One of the advantages of these methods is that they explore the causal nature of gene regulation rather than simply gene-gene association.

Because the majority of cancer data does not offer explicit information about dynamic profiles of gene expression, most of the current methods used for causal inference in cancer research are developed for static data. Thus, these methods overlook the fact that expression levels of certain genes change throughout the

cell cycle. Therefore, they do not characterise the regulatory relationships that occur *during* cancer development. Furthermore, the majority of current methods are based on strong assumptions that contradict heterogeneous data, such as real cancer data [60].

Causal inference can be performed using different approaches, and thus we have subdivided this category of methods into *Granger causality*-based methods, *causal Bayesian* methods, and other causality methods.

3.1.4.1 Granger causality

Granger causality (GC), which is widely used in economic studies, is starting to indicate promising results in the biological and medical fields [118]. The idea is the following: given two time-series vectors **Y** and **X**, **Y** can be defined as the *Granger cause* of **X** if **Y** can assist in predicting the future of **X**. In other words, if past observations of **Y** assist in reducing the prediction error for **X** (when compared to using only past observations of **X**), **Y** is considered to *Granger-cause* **X** [119, 118].

Because *Granger causality* assesses the utility of one time series in predicting another, it does not necessarily represent real regulatory relationships or describe the dynamics of the inferred relationships. Furthermore, this approach may produce misleading results when the true regulatory relationship involves more than two genes. This fact limits the usability of *Granger causality*-based approaches used in cancer research, as most of the regulatory interactions that occur during the progression of the disease involve more than two genes. However, these methods are capable of identifying parts of the GRNs when gene expression time series data are provided.

An example of a GRN method that employs GC is described in [118]. In this method, a pairwise comparison used to detect GC is carried out through the bivariate autoregressive (AR) model, which is expressed as

$$\begin{aligned} x_{1,t} &= \sum_{l=1}^{p} (a_{11,l} x_{1,t-l} + a_{12,l} x_{2,t-l}) + e_{1,t}, \\ x_{2,t} &= \sum_{l=1}^{p} (a_{21,l} x_{1,t-l} + a_{22,l} x_{2,t-l}) + e_{2,t}, \end{aligned} \tag{3.5}$$

where p corresponds to the model order (number of time lags), $a_{ij,l}$ are the coefficient values of the AR model, and $e_{1,t}$ and $e_{2,t}$ are the residuals (prediction errors) of time series 1 and 2, respectively. The method tests the following null hypothesis: $H_0 = j$ *does not Granger-cause variable* i. The gene j Granger-causes the gene i if a small *p-value* is obtained while the hypothesis is tested.

CGC-2SPR [120], which is based on *Granger causality*, is a method that uses time series data to model GRNs. It attempts to reduce the false positive rates for datasets in which the number of time series (corresponding to genes) is significantly higher than the number of data points (this is common for gene expression time series). To achieve this goal, CGC-2SPR consider all genes together by using *Ridge regularisation* and by incorporating a prior knowledge weight matrix W with pairwise gene associations.

Regulatory relationships are defined by testing the same null hypothesis ($H_0 = j$ *does not Granger-cause variable* i) and conserving the links for which the p-value of the hypothesis test is small enough to reject the null hypothesis.

The *Prophetic Granger Causality* **(PGC)** method [121] was developed to infer GRNs from protein-level time series data. This method receives as input a dataset $\mathbf{X}$ of n time series that contain information on the levels of a set of m proteins. PGC uses the *LASSO-Granger* method described in [122], according to which the predicted level of a protein p at time instant t in the time series i ($\hat{y}_{i,p,t}$) can be expressed as a linear combination of the present and past levels of the other observed levels contained in $\mathbf{X}$.

PGC introduces the concept of *prophetic extension*, in which the regression can incorporate information on the future states of other proteins. Given two variables p and q, the future states of p can (at least partially) provide information that allows the future behaviour of q to be predicted. This method also incorporates prior knowledge to improve its performance. A prior knowledge matrix B is computed using a heat diffusion approach that employs only interactions from the literature and does not refer to the time series datasets.

SINCERITIES (SINgle CEll Regularized Inference using TIme-stamped Expression profileS) [123] infers regulatory relationships between TFs and regulated genes.

The method quantifies the temporal changes in the expression of each individual gene as the *Kolmogorov–Smirnov* (KS) distance between two subsequent points in time. SINCERITIES uses the distance vectors as the target of a *Ridge* regression model. The inferred coefficients of the regression and a partial correlation analysis are used to define the strength of the regulatory edges. Additionally, *Granger causality* is applied to the distance vectors to determine the direction of the regulation.

Finally, one of the methods in this category that has become popular in the last years is **Scribe** [124], which performs GRN inference using a combination of *Granger causality* and *mutual information*. The method implements the concept of *restricted directed information* (RDI) to infer the mutual information shared between the immediate past of a putative regulator and the present of a target gene. The direction of the relationship is assigned following the logic of *Granger causality*. The method aims to control the false positive rate by conditioning the mutual information on all genes other than the analysed regulator-regulated pair.

3.1.4.2 Bayesian network based methods

The causal Bayesian network method describes the nature of interactions between network elements as a *directed acyclic graph* (DAG), in which the nodes represent random variables $\mathbf{X} = X_1, X_2, ..., X_p$. In the specific case of GRNs, the node X_i corresponds to the expression of gene i, and the edges connecting the different nodes in this network represent functions of the conditional distribution of a node given its parents. In general, this distribution can be either multinomial or continuous, depending on whether the variables are discrete or continuous, respectively.

In these methods, the whole structure of the network at one particular moment (denoted as $\mathbf{x}$) can be described as it is in [105]; the discrete and continuous cases are

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{p} \mathbb{P}(X_i = x_i \mid \mathbf{X}_{pa(i)} = \mathbf{x}_{pa(i)}), \tag{3.6}$$

$$f_{\mathbf{x}}(\mathbf{X}) = \prod_{i=1}^{p} f_{x_i}(x_i \mid \mathbf{X}_{pa(i)} = \mathbf{x}_{pa(i)}), \tag{3.7}$$

respectively, where $\mathbf{X}_{pa(i)}$ represents the subset of genes acting as parents of X_i, and $\mathbf{x}_{pa(i)}$ represents their respective values at that particular moment.

In 2013, Bayesian network theory was applied to causal inference in cancer biology for the first time [125]. The method employs Bayesian network learning and graph search techniques to construct a comprehensive three-component regulatory network (miRNA-TF-mRNA) by integrating data from various sources. The initial network structure is established using target information that captures miRNA-mRNA, miRNA-TF, TF-miRNA, TF-TF, and TF-mRNA interactions. This method uses bulk gene expression data from different biological conditions (sample categories) to infer a Bayesian network structure for each biological condition. The method assembles the resulting networks into an integrated global network by implementing a bootstrapping and averaging procedure that retrieves the interplay between miRNAs and TFs and the network motifs that involve at least two regulators.

Another classical method that implements causal Bayesian networks is Bayesian Network Inference with Java Objects **BANJO** [126] (and its variations, such as BANJO-weighted symbolic model checking [127]). **BANJO** employs a first-order Markov dynamic Bayesian network (DBN) in which each node is influenced by itself and its parents' values in the immediately preceding instant in time. This method evaluates how well different networks explain the behaviour of observational data using a scoring metric based on the full Bayesian posterior probability $P(G \mid D)$ for each created network G given the observed data D.

Similarly, **CRNET** [128] was proposed as an integrative method that employs previous knowledge about the activated form of TFs to model the gene expression $\mathbf{y_i}$ as

$$\mathbf{y}_i = \sum_{t=1}^{\mathcal{T}} z_{j,t} a_{j,t} \boldsymbol{x}_t + \eta_j + \mathbf{n}_j, \tag{3.8}$$

where $\mathbf{x}_t = [x_{t,1}, ..., x_{t,m}, ..., x_{t,M}]^T$, $x_{t,m}$ is the t^{th} TF in the m^{th} sample, $z_{j,t}$ is the binding state ($z_{j,t} = 0$ when unbounded, $z_{j,t} = 1$ when bounded), $a_{j,t}$ is the binding strength ($a_{j,t} = 0$ when $z_{j,t}$=0), η_j represents the basal expression at $t = 0$, and $\mathbf{n}_j$ is a Gaussian additive noise function with a mean of zero and a variance denoted as σ^2.

Unlike other methods, the problem with **CRNET** is not optimisation, but the inference of hidden TFs through a iterative two-stage Gibbs sampling procedure. With this approach, functional bindings can be assessed by testing the following

null hypothesis: H_0 = *no binding (regulation strength equals zero)*. The performance was tested according to five benchmark regulatory networks used in the DREAM4 challenge, and the results suggest that including previous knowledge of TF activities improves the convergence, rendering it faster than similar Bayesian approaches.

GRNVBEM is a GNR inference method that employs a Bayesian framework to fit time series gene expression data via a first-order autoregressive moving average (ARMA) [129]. In this method, given a set of G genes, the GRN can be described by employing two features:

- Linkage: $\mathbf{X_i} = [x_i(1), \ldots, x_i(j), \ldots, x_i(G)]^T$, where $x_i(j) = 1$ means that the j^{th} gene is the parent of the i^{th} gene, and $x_i(j) = 0$ otherwise.

- Regulatory effect: $\boldsymbol{\omega}_i = [\omega_i(1), \ldots, \omega_i(j), \ldots, \omega_i(G)]^T$, where $\omega_i(j) > 0$ represents activation and $\omega_i(j) < 0$ represents inhibition.

In this method, GNR inference is performed using a dataset Y with G genes and $N+1$ samples, where $y_i(n)$ is the fold change of the i^{th} gene at the n^{th} point in time. The model is approximated by a first-order autoregressive (AR) model. Therefore, the fold change of gene i at time instant n (an unknown value denoted as $z_i(n)$) is a linear combination of its parents at the immediately preceding point in time $(n-1)$ and is described by

$$y_i(n) = \sum_{j=1}^{G} z_j(n-1)\omega_i(j)x_i(j) + e_i(j), \tag{3.9}$$

where $e_i(j)$ represents the error of the current time sample and is modelled as an independent and identically distributed Gaussian with a mean of zero and unknown variance, which is expressed as

$$p(e_i(n)) = \mathcal{N}(e_i(n) \mid 0, \sigma_i^2), \forall n. \tag{3.10}$$

Because of the fact that $z_i(n)$ can be expressed as the difference between the data and the noise (i.e. $z_i(n) = y_i(n) - e_i(n)$), the inference problem can be rewritten as an optimisation problem of the posterior $p(\mathbf{x_i}, \boldsymbol{\theta}_i \mid \mathbf{Y})$ in terms of the likelihood function and the prior [129] according to:

$$y_i(n) = \sum_{j=1}^{G} (y_i(n-1) - e_j(n-1))\omega_i(j)x_i(j) + e_i, \tag{3.11}$$

$$p(\mathbf{x_i}, \boldsymbol{\theta}_i \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{x_i}, \boldsymbol{\theta}_i)p(\mathbf{x_i}, \boldsymbol{\theta_i})}{p(\mathbf{Y})}. \tag{3.12}$$

3.1.4.3 Other causal methods

Several methods that utilise causal approaches other than *Bayesian* and *Granger causality* for GNR inference are now described.

One of the pioneering methods of applying causal theory to biological problems was developed by Le et al. in 2013 [130]. In their method, miRNA-RNA regulatory relationships were inferred from gene expression data. To accomplish this, the causal effects between genes were deduced by using a *completed partially directed acyclic graph (CPDAG)*. To construct the CPDAG, the method involved testing the conditional independence of each miRNA-mRNA pair to infer a CPDAG by conditioning the miRNA-mRNA pair to different sets of genes (nodes) according to the iterative process known as the *PC algorithm* [54]. Finally, the method determined the strength of the causal link (causal effect) via *do-calculus* [131].

A recently developed example is **mCADET** [128], which is a hybrid method useful for discovering GRNs in cancer through miRNA-mRNA causal relationships. In this method, a deep learning algorithm is used to simulate the transcription mechanism based on the correlation between miRNA-mRNA pairs.

This method is based on the concept of hidden regulators co-regulating the miRNA-mRNA pairs. A neural network is trained to find an architecture that explains highly correlated miRNA-mRNA pairs using a co-regulation mechanism. However, a high level of correlation does not necessarily imply causality. Thus, mCADET includes information on confirmed regulatory relationships to allow the deep learning algorithm to estimate the causality and direction of the regulation in the inferred network through a deep belief network (DBN) model that contains directed edges from the miRNA.

The Fused-Sparse Structural Equation Model **(FSSEM)** [103] is an algorithm based on a structural equation model (SEM) framework. **(FSSEM)** infers causal rela-

tionships and constructs GRNs under two different conditions ($k = 1, 2$) to later identify differences between the inferred networks. The method assumes the existence of an observed set $\mathbf{x}_i^{(k)}$ of genetic perturbations that occur naturally, which are defined as those originated by different genotypes of expression quantitative trait loci (eQTLs). The observed q *cis*-eQTLs can be formalized as

$$\mathbf{x}_i^{(k)} = [x_{i1}^{(k)}, x_{i2}^{(k)}, ..., x_{iq}^{(k)}]^T. \tag{3.13}$$

The method also requires information on n gene expression levels from two distinct conditions (e.g. micro-array and RNA-Seq), which are denoted as

$$\mathbf{y}_i^{(k)} = [y_{i1}^{(k)}, y_{i2}^{(k)}, ..., y_{in}^{(k)}]^T. \tag{3.14}$$

In both cases, $i = 1, 2, ..., n_k$ indexes the number of individuals. Taking into account the fact that gene expression can be regulated by other genes and may be influenced by its eQTLs, the proposed SEM model was

$$\mathbf{y}_i^{(k)} = \mathbf{B}^{(k)}\mathbf{y}_i^{(k)} + \mathbf{F}^{(k)}\mathbf{x}_i^{(k)} + \boldsymbol{\mu}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \tag{3.15}$$

where $\mathbf{B}^{(k)}$ is an $n \times n$ matrix that defines the unknown network structure under condition k, $\mathbf{F}^{(k)}$ models the effect of *cis*-eQTLs under condition k on gene expression, $\mu^{(k)}$ is the SEM bias, and $\epsilon^{(k)}$ represents the residual error in the model. The FSSEM algorithm adopts the proximal alternating linearised minimisation (PALM) method [132] to solve the problem; the implementation details can be found in [103].

mirTime [102] is another recently developed method that uses previous knowledge in order to identify miRNA-RNA regulatory relationships. This method claims causality owing to the fact that miRNA-RNA relationships are always down-regulatory, with the miRNA as the parent. In **mirTime**, the dynamics are approximated using a Gaussian process (GP) model for the purpose of calibrating time series gene expression data and synchronising target-predictor datasets. Candidate targets were chosen as the intersection between the following items:

- Genes predicted in the TargetScan 7.2 databases [133]
- miRDB 5.0 [134]
- MirTarget 3.0 [135]
- Genes from the dataset predicted using TargetExpress [136]
- Genes experimentally confirmed as targets in miRTarBase 7.0 [137]

mirTime uses an approximated model to obtain extra time data points so that all predictor-target data match and can be clustered via a spherical k-means algorithm. Each cluster contains a subset of target genes that can be targeted by an mRNA. Relationships are evaluated by a correlation-based to remove inconsistent links.

DeepSEM adopts a beta-variational autoencoder through a SEM that infers causal relationships using a linear model [138]. In this method, the optimised weights of both the encoder and decoder functions are assumed to represent the adjacency matrix of the GRN. One of the main drawbacks of this method is that the running time increases with the number of genes involved due to the "inverse" operation in the inverse GRN layer. Consequently, its applicability to real cancer data is limited.

3.2 Methods for discovering cancer drivers

It has been theorised that cancer is the physical manifestation of a relevant number of causal somatic mutations that appear sequentially, starting with a mutation that provides a selective advantage to the cell that initiates tumorigenesis [139]. A growing area of research is focused on identifying cancer drivers, which is a challenging task because there are a large number of mutations (called "background" or "passenger" mutations) that do not lead to cancer.

The complete identification and understanding of cancer-causing genes is critical in cancer research because of the significant impact that such knowledge has on the development of diagnostic methods and effective treatments. For this reason,

a large number of computational methods that attempt to identify cancer drivers have been developed.

Most of the widely accepted methods use mutational information to determine the significance, functional impact, or structural consequences of mutations [140]. The rationale behind these approaches is based on the hypothesis that cancer drivers are mutated genes. A number of other approaches have incorporated biological network information to improve predictions of cancer drivers. The idea behind network approaches is that frequently mutating central genes in biological networks play a more significant role in cancer than those that mutate less frequently.

Classical biological research on the genes that drive cancer underestimates the complex dynamics involving such genes *during* the progression of the disease. This is the case because cancer driver studies are strongly biased towards the detection of gene mutations, and they neglect the importance of the dynamic alterations of the processes that cause the disease. Tumorigenesis and cancer progression are complex and heterogeneous processes that cannot be easily described in terms of the frequencies of driver and passenger mutations. Despite providing valuable information, mutation- and network-based methods are not capable of representing the transformation of a normal gene into a cancer driver gene.

Furthermore, current methods tend to focus on analysing probability distributions of population data without considering time as a relevant factor. Such probability distributions usually reflect the landscape at advanced (also known as *steady-state*) stages of the disease. Hence, these methods fail to gather valuable information about the temporal behaviour of tumour progression.

A crucial point frequently overlooked in current methods is that all cancer driver genes dysregulate critical biological processes in the cell. This dysregulation can be induced by highly mutated genes or genes without any mutations at all. A strong implication of this fact is that cancer-causing genes are not fully characterised by the number of mutations they exhibit, nor by the number of connections they have in GRNs. The key feature of cancer drivers is the fact that their genetic aberrations affect the dynamics of biological processes in such a way that they induce carcinogenesis and tumour progression.

Therefore, in this thesis, cancer drivers are linked to biological processes with the altered dynamics that cause cancer to develop. Identifying the genes that cause alterations in the dynamics of biological processes *during* cancer progression is critical for a complete understanding of the disease.

In the following subsections, the main categories, primary characteristics, and current limitations of prominent methods used for discovering cancer drivers are summarised.

3.2.1 Mutation-based methods

3.2.1.1 Frequency-based methods

The most widely accepted method for detecting cancer-causing genes is to identify genes with a significantly higher number of somatic mutations than that expected as a result of random chance. The first approaches assumed that the background mutation rate was constant and that it was shared among all samples [141]. However, this is not the actual case. It is now well known that the mutation rate can be influenced by different features such as the mutation type, base at the mutation site, nucleotide context, and bases neighbouring the mutation (dinucleotide context).

These findings have been used to improve the detection of frequency-based methods. For example, **MuSiC** [142] is an algorithm that allows different mutation rates for different mutation types. This formed the basis for **MutSigCV** [143], which included additional factors such as expression level and replication timing. These methods evaluate the significance of mutations in DNA sequencing to discover cancer drivers.

WeSME [144] is a method that accounts for the fact that several mutations in cells are mutually exclusive. It computes the mutation frequencies of patients from the observed distribution, and then uses a permutation test to approximate the result of a pairwise mutual exclusivity estimation. **WeSME** performs this approximation without the high computational cost required for rigorous p-values. It effectively delivers mutual exclusivity networks in which connected nodes correspond to mutually exclusive genes.

Cancer drivers inferred by frequency-based methods are genes that exhibit somatic mutations at a higher rate than that expected by random chance. The logic behind this idea (which is widely accepted even now) is that highly mutated genes are more likely to cause cancer. However, frequently mutating genes do not necessarily transform into cancer drivers. Certain genes can harbour a large number of passenger mutations that do not contribute to cancer development. Thus, such genes are not necessarily cancer drivers.

3.2.1.2 Hotspot-based methods

OncodriveCLUST [145] is one of the most benchmarked methods for identifying mutations that change one of more aspects of a cell's phenotype (commonly known as non-silent mutations). This method is capable of predicting the harbouring of mutation clusters in protein sequences more frequently than expected based on the distribution of synonymous background mutations. The basic steps involved in OncodriveCLUST are as follows:

1. Retrieve single-nucleotide protein affecting mutations.

2. Using a binomial cumulative distribution function that accounts for the gene length and overall number of gene mutations, determine the positions that have a probability of a mutation occurring of greater than 1%.

3. Create clusters of those positions such that positions within five amino acids form a single group.

4. Include all other mutation positions not used during the binomial analysis that are in or are adjacent to the cluster.

5. Score each cluster as follows: Clustering Score $= \sum_i \frac{FM}{(\sqrt{2})^{dist}}$, where i is the protein position, FM is the fraction of mutations in that protein over all samples, and $dist$ represents the number of amino acids between i and the position of the cluster with the largest number of mutations.

OncodriveCLUST is based on mutation clustering, and thus rare cancer driver mutations are invisible to the method. However, the method has proved to be useful in complementing the results of other cancer driver inference methods.

MutSig2CV (2014) [146] identifies genes that have mutated more frequently than expected relative to a certain background mutation rate (BMR). Thus, local sites with a significant number of mutations are considered hotspots and are clustered. The inference process is complemented by using evolutionary conservation information to determine the functional significance of a mutated site, assuming that places conserved across vertebrates are more significant than other places in the genome.

MSEA [147] is a similar method that uses the KS test to identify clustering patterns of mutations. The method defines a "mutation set enrichment analysis" (MSEA) to identify mutation hotspots. In these hotspots, genes that have a mutation rate higher than the rest of the domain are classified as cancer drivers.

MutSpot is a recently developed method that scans the entire genome for mutation hotspots by fitting a genomic background model. The method also allows for a local search in "regional discovery mode". In this mode, MutSpot fits a background model to user-defined regions such as promoters, and this enables the method to identify hotspots in the specified local region [30].

MutSpot's hotspot detection method uses a LASSO logistic regression model to estimate patient- and position-specific background mutation rates. The method corrects for known covariates of mutation probability, such as local nucleotide context, replication timing, and epigenomic features. MutSpot searches throughout the genome for mutation recurrences within *l-base pair* regions with at least n mutated samples (default $l = 21$, $n = 2$). To account for changes in mutation rates across different patient tumours, a Poisson binomial model is used for the p-value when computing the mutation recurrences.

Hotspot-based methods are widely recognised for their value because the recurrence of somatic mutations (which are hotspots by definition) is considered universal evidence of somatic positive selection. However, mutation hotspots are not necessarily linked to cancer drivers. In a recent study, the hypothesis that hotspots unequivocally pinpoint genes that drive tumorigenesis was shown to be incorrect because it is based on an inaccurate statistical model of background mutagenesis [148]. In fact, the same study showed that many mutation hotspots are in fact pas-

sengers. Therefore, recurring mutations at inherently mutable sites may not imply positive selection.

3.2.1.3 Functional impact-based methods

A natural solution to the problem of bias in frequency-based approximations is to review the "functional impact" of mutations. In this approach, the frequency of mutations is important in the detection of cancer drivers if such mutations appear in genes that have a substantial impact on cellular functions. From this point of view, a mutated gene with a significant impact on biological processes is more likely to be a cancer driver because of its ability to alter, for example, various aspects of the cell phenotype.

One of the benchmarks of this class of methods is **OncodriveFM** [149], which analyses functional mutations to detect the pathways and genes relevant to cancer. This method detects and prioritises the pathways and genes that exhibit a bias towards the accumulation of functional somatic variants. In this context, genes that show such a bias are more likely to be cancer drivers. In addition, OncodriveFM can detect cancer drivers harbouring a low number of mutations because the bias analysis does not rely on the frequency of the somatic mutations.

DriverML [150] is a recent method that integrates mutational frequency and functional impact information. This machine learning method infers cancer driver genes by using *Rao's score* to create a weighted global statistic that considers each type of mutation (i.e. missense, nonsense, splice site, frameshift insertion/deletion, and in-frame insertion/deletion). A background mutation model is used to find optimal weights. The score statistics found after training quantifies the functional impacts of different types of mutations. A *Monte Carlo* procedure is then used to determine the genes that have statistically significant scores according to real data (i.e. scores that are significantly greater than simulated scores).

The **cTaG** method [151] also implements the idea of combining frequency and functional impact information to discover cancer drivers. This method is a radiometric approach (i.e. it uses the relative proportions of different types of mutations) to identify tumour suppressor genes (TSGs) and oncogenes (OGs). The method implements a supervised ensemble learning approach to classify genes as

TSG, OG, or neither based on 19 features, including frequency of different mutations, type of mutation, and functional impacts. The more classifiers that identify a gene with the same label (TSG or OG), the higher the confidence that the gene is a cancer driver.

In general, functional impact-based methods are suitable for detecting cancer drivers with low mutation rates. However, these approaches provide minimal information on the behaviour of inferred cancer drivers, especially during the early stages of the disease. The functional impact of a mutation is usually detectable at advanced stages of cancer progression. Moreover, functional impact-based approaches do not describe how gene interactions are affected throughout the development of the disease. Thus, the utility of this type of method is limited.

3.2.2 Network-based methods

Network-based methods acknowledge the critical fact that a disease phenotype is frequently a consequence of alterations in biological processes [152]. This acknowledgement is based on the fact that biological processes are the result of molecular interactions that can be represented as networks. Accordingly, network-based methods employ prior knowledge from biological networks to evaluate the significance of specific genes and their potential link to tumorigenesis. The rationale of these methods is that using previous knowledge about interactions between genes and proteins enhances the prioritisation of putative drivers [153].

TimeTP [154] analyses time series data to detect perturbed sub-pathways by creating an augmented graph of the networks. It also infers TFs that trigger perturbations in the networks. In this method, an augmented node corresponds to a gene containing a time vector $\vec{v}$, for which the value at one point in time can be -1 (underexpressed), 1 (overexpressed), or 0 (no change). This vector is used to infer whether or not the edges in a perturbed sub-pathway propagate the differential expression pattern ($\vec{v}$).

Specifically, for any 2 genes G_1 and G_2 with time vectors $\vec{v}_1$ and $\vec{v}_2$ (respectively), if $\vec{v}_1$ is propagated from G_1 to G_2, then $\vec{v}_2$ should be similar to $\vec{v}_1$. The cross-correlation $(\vec{v}_1 \star \vec{v}_2)(n)$ was used as the similitude measurement. Once edges that

transmit their temporal patterns are found, the delay time is calculated as the delay for which the cross-correlation is maximised. This information allows TimeTP to infer TFs that trigger those perturbations as possible cancer drivers.

Despite using time series data, this method provides few insights into the real dynamics of gene expression, as the data is discretised using only three levels. Furthermore, this method uses similarity measurements rather than causality tests to identify cancer drivers. Therefore, TimeTP neglects the causal aspect of a cancer driver.

In 2016, Vinayagam et al. [155] proposed a method that uses engineering control theory to identify genes that are relevant to biological networks. The method hypothesises that relevant genes detected in a protein-protein interaction (PPI) network are promising candidates for cancer drivers. Vinayagam et al. conducted a controllability analysis on the PPI network to define a *minimum driver node set* (MDS), which is the smallest set of genes required to "control" the network (in the sense of engineering control theory).

A controllability analysis is performed by applying control theory to the matrix representation of the network. The MDS is further used to classify the nodes as "indispensable", "dispensable", or "neutral". A node is considered "indispensable" if its removal from the network increases the size of the MDS (inferred from the new synthetic network). Finally, genes that change their indispensability when the network is perturbed (e.g. through connection rewiring or edge flipping) are identified as putative cancer drivers.

CBNA, a more recent method developed by Pham et al. in 2019 [75], uses aspects of control theory to discover cancer drivers. **CBNA** can identify coding and non-coding cancer drivers by detecting "critical nodes" in a network with miRNAs, TFs, and mRNAs as nodes. The network structure is inferred by combining similarity measurements (correlation) and previous biological knowledge. A matrix representation of such a network is further used to test Kalman's controllability condition, which is widely accepted in modern control theory. The main goal of **CBNA** is to find a *controllability* matrix **C** that contains the minimum set of nodes (cancer drivers) that can "control" the network in the sense of Kalman's controllability condition [156].

Even though the controllability concept is strongly based on the state-space representation of a system, which is an inherent representation of dynamic systems, the methods developed by Vinayagam et al. and those developed by Pham et al. do not include a temporal aspect in the data. In fact, control theory (both classical and modern) deals with the control of dynamic systems. Therefore, the controllability analyses implemented in current methods used for inferring cancer drivers are, at best, focused on the steady-state behaviour of GRNs.

Another example of a recent development based on biological network analysis is **PRODIGY** [157]. PRODIGY is a method used for ranking cancer drivers in single-patient data, and thus it is applicable for inferring personal cancer drivers. The method assumes that cancer driver mutations influence the deregulation of other genes because of their high connectivity within the affected pathways. PRODIGY infers and scores cancer driver genes based on their connections by using a variant of the prize-collecting Steiner tree algorithm.

In this method, a sub-network within a weighted graph is found as a result of an optimisation process that maximises the score of a mutated gene g that belongs to the set of nodes V_T (i.e. $g \in V_T$) of a network defined by a set of nodes V and a set of edges E. Scoring is performed as shown below:

$$Score(T) = \sum_{v \in V_T} P(v) - \sum_{(u,v) \in E_T} W(u,v), \tag{3.16}$$

where $W : E \rightarrow \mathbb{R}_+$ is a positive weight function on the edges and $P : V \rightarrow \mathbb{R}$ is a weight function on the nodes.

In **PRODIGY**, nodes acquire a positive value if the corresponding gene is a differentially expressed gene (DEG) between healthy and cancerous tissue (from the same patient and the same kind of tissue). A negative value is assigned to a node if the gene that it represents is not differentially expressed between healthy and cancerous conditions. The inferred networks in PRODIGY assume edges as penalties. Therefore, the final score of a sub-tree T is calculated as the sum of all rewards (nodes) and penalties (edges), always considering that the root of the tree is a mutated gene.

Accordingly, every mutated gene can be assessed to provide a landscape of the importance of such genes as cancer drivers. Additionally, the method uses the obtained sub-tree as an indicator of how that cancer driver deregulates the network and affected pathways.

sysSVM [158] is a machine learning model that implements a support vector machine to determine cancer drivers at the single-patient level. This method hypothesises that somatic alterations that transform a gene into a cancer driver affects genes with specific properties. The properties analysed include genomic, epigenomic, gene expression, protein domain, and network features. This approach employs these properties rather than the mutation frequency to identify cancer drivers.

In addition, sysSVM uses the support vector machine approach to identify genes that have properties similar to well-known cancer drivers. The strength of a gene's contribution to cancer is assumed to be closely linked to the similarity between the gene and known cancer drivers. The method uses this similarity to identify cancer driver genes.

sysSVM2 [159] is a recent optimisation of the sysSVM method that incorporates cancer genetic alterations and previous knowledge from different biological networks. This method removes highly correlated features that may compromise the performance of the support vector machine. Furthermore, the method implements an additional optimisation procedure for hyper-parameter tuning.

Although network-based methods can successfully address a number of the limitations inherent to mutation-based approaches, they are strongly biased towards steady-state representation of the altered processes involved in cancer. Consequently, the current methods in this category are not suitable for modelling the dynamics of gene interactions *during* cancer.

3.2.3 Causality based identification of cancer drivers

This category of methods for inferring cancer drivers implements formal causal approaches such as *causal Bayesian* inference to determine whether or not a gene is a cancer driver.

CAPRESE [160] employs the idea of "oncogenetic trees" to define causality based on ecological models of cancer progression. An oncogenetic tree usually models cancer development as independent acyclic paths forming a tree structure. In these trees, nodes represent important genetic events, and arcs or edges imply transformations from a previous state to another in strict sequential order.

CAPRESE is based on the notion of *probabilistic causation* under the framework proposed by Suppes [161], according to which, given two events A and B, it is possible to infer causality if:

- A occurs before B
- The occurrence of A increases the probability that B can be observed

Formally, this means that A can be considered a *prima facie* cause of B, as claimed in [160], if

$$P(B \mid A) > P(B \mid \overline{A}) \; and \; P(A) > P(B), \tag{3.17}$$

where $P(A) > P(B)$ is an indicator of temporal precedence when considering the probability raising (PR) property in this framework. However, this basic notion of causality is not able to address certain problems regarding the determination of genuine causes. As a result, a filtering process is also required to reduce the number of spurious causes.

CAPRI [91] follows Suppes' framework, but incorporates formulae that use Boolean connectives to identify progression patterns, which are defined as "specific combinations of the closest upstream events that confer a selective advantage". In **CAPRI**, it is assumed that, for an event e (e.g. the appearance of a mutation), a *conjunctive normal form* formula φ (a Boolean formula in terms of "and" $\wedge$, "or" $\vee$, and "xor" $\otimes$) can describe the pattern and the condition for selectivity.

Following this definition, φ can be formulated as

$$\varphi \rhd e \Leftrightarrow P(\varphi) > P(e) \; and \; [P(e \mid \varphi) > P(e \mid \overline{\varphi})], \tag{3.18}$$

where $\rhd$ indicates a causal relationship and a single φ for mutations in genes a, b, c can be either $a \vee b \rhd c$, $a \otimes b \rhd c$, or $a \wedge b \rhd c$, which are understood as $a \rhd c$ and $b \rhd c$.

These two methods describe only the temporal patterns of the appearance of mutations, and thus they do not describe the dynamics of the gene interactions or the implications for the biological processes that cause tumorigenesis. Moreover, temporal precedence is necessary but not sufficient for determining real causality. Therefore, "probabilistic causation" as defined by these two methods is prone to spurious findings.

MADGiC [162] is a unified empirical Bayesian model-based approach for identifying cancer driver genes. Based on the idea that several features can be integrated into a unified method, MADGiC uses the mutation frequency, evidence for functional impact, and spatial pattern of mutations to optimise the detection of cancer drivers.

This method defines a set of G genes (indexed by g) with at least one non-silent somatic mutation (i.e. a mutation that causes a change in the amino acid sequence) as well as a set of J samples (indexed by j) with at least one non-silent somatic mutation per sample in one or more of the G genes. These sets are further used to create the observed non-silent mutation states vector, functional impact scores vector, and driver status flag:

- $\vec{X_g} = X_{1g}, X_{2g}, ..., X_{Jg}$ is the binary vector of observed non-silent mutation states, where "1" means that one or more non-silent mutations are present anywhere in the gene g.

- $\vec{S_g} = S_{1g}, S_{2g}, ..., S_{Jg}$ is the vector of functional impact scores for non-silent mutations in gene g for all samples using Sorting Intolerant From Tolerant (SIFT), a method that predicts if protein function is affected by an amino acid substitution.

- $Z_g \in 0, 1$ is the indicator that gene g exhibits driver activity.

MADGiC uses a *probabilistic generative model* to calculate the posterior probability of a gene being a cancer driver. Given the mutation status and impact score of a gene across J independent samples, such a gene is a cancer driver if its posterior probability of being a cancer driver is greater than 0.95.

The **Tumour-specific Causal Inference (TCI)** [163] method uses causal Bayesian inference to identify cancer drivers. In TCI, given a tumour t, with its respective sets of somatic genome alterations SGA_SET_t and differentially expressed genes DEG_SET_t, causal relationships are evaluated by using a bipartite causal Bayesian network. The method searches for the tumour-specific causal model M_t by optimising the posterior probability $P(M_t \mid D)$, in which the data D contains the two sets described previously. In this approach, D corresponds to a collection of genomic data from The Cancer Genome Atlas (TCGA) Research Network (https://www.cancer.gov/tcga).

Causality in the method is inferred using the formal Bayesian framework expressed as

$$P(A_h \rightarrow E_i \mid D) = \frac{1}{Z} P(A_h \rightarrow E_i) P(D \mid A_h \rightarrow E_i), \tag{3.19}$$

where E_i represents the i^{th} DEG event in DEG_SET_t, A_h is the most likely SGA belonging to SGA_SET_t that is a causal predictor of E_i (denoted as $A_h \rightarrow E_i$), and Z is a normalisation term calculated as

$$Z = \sum_{j=0}^{|SGA_s ET_t|} P(A_j \rightarrow E_i) P(D | A_j \rightarrow E_i). \tag{3.20}$$

These two equations rely on calculating the prior probability that A_h causes E_i (namely $P(A_h \rightarrow E_i)$) and the conditional probability of data D given that $P(A_h \rightarrow E_i)$. This allows the method to evaluate how well A_h can explain the variance of E_i in tumours hosting A_h.

3.3 Gene Group Analysis of Cancer Drivers

The dysregulation of biological processes that lead to cancer is not the consequence of alterations in a single cancer driver gene. For carcinogenesis and tumour progression to occur, it is necessary for multi-gene systems to harbour genetic aberrations that cause the dysregulation of key processes in the cell. Recent experiments have shown that biological processes can be controlled through the regulation of the joint activity of groups of genes (an example can be found in [164]).

Gene group analysis has emerged as a new cancer research area in recent years. Gene group analysis typically involves the identification of sets of genes that are

associated with a particular phenotype or state in cancer progression. The primary aim is to clarify the multi-gene mechanisms that govern cancer progression. The rationale behind gene group studies is the fact that cancer-causing genetic alterations can be observed more frequently at the level of macroscopic biological processes than at the nucleotide or gene level [73].

Although gene group-based methods consider the joint effect of sets of genes on cancer, they may not fully explain the molecular mechanisms that underly cancer development. Specifically, current methods neglect the dynamics of the interactions between genes that are involved in the progression of the disease. Current gene group analysis methods focus on the joint effects observed at a single snapshot in time. Hence, they frequently overlook the fact that gene-phenotype associations change throughout the progression of cancer. Therefore, the set of cancer-causing genes can differ depending on the stage of the disease.

Many gene modules, for instance, are built using co-expression networks that use the correlation between the gene expression levels in a single sample or a limited number of samples. These networks may distort the underlying biological interactions between the genes involved in cancer because gene regulation is frequently impacted by genetic alterations that vary throughout cancer progression. Furthermore, co-expression networks are not necessarily causal, and thus the causal aspect of the disease is overlooked in these methods.

Multi-gene association is typically described in terms of *gene modules*. A gene module is a group of functionally related genes that are co-expressed or co-regulated in cancer cells. Multi-gene modules are typically linked to network representations of the biological processes that are altered in cancer (e.g. cell proliferation) as well as the signalling pathways that are activated or dysregulated in cancer cells.

One early example of a gene group-based method is **MEMo** [165]. This method identifies sets of recurrently altered genes that are likely to belong to the same pathway or biological process. Genes identified by MEMo are restricted to significantly mutated genes that exhibit mutually exclusive genetic alteration patterns across multiple patients. The method hypothesises that modules formed by genes with mutually exclusively genetic alterations are likely to drive cancer progression.

iMCMC [166] constructs a gene expression-based network in which the nodes are genes and the edges are gene expression correlations (called the *expression network*). In iMCMC, an additional network with the same topological structure as the one inferred from the gene expression is created. In the second network (namely, the *mutation network*), nodes are weighted to reflect the frequency of the mutations of the genes across all samples. The edges in the *mutation network* represent the number of all samples in which exactly one of the linked genes is mutated divided by the number of samples in which at least one gene in the pair is mutated.

A final network is formed by combining the *expression network* and the *mutation network*. Gene modules detected by iMCMC correspond to the genes that form sub-networks with "high weights" in both edges and vertices. Such sub-networks are identified by following the optimisation model described in [167]. The method hypothesises that the joint dysregulation of the identified multi-gene modules drives cancer progression.

The network-based approach **HotNet2** [168] represents genes as the nodes and interactions between the encoded proteins as the edges. The method employs a PPI network as prior knowledge in order to find small sub-networks in which mutations occur at a rate higher than that expected by chance. HotNet2 uses a heat-diffusion process model to infer a data-specific network. The diffusion model uses as input a heat vector containing a score (e.g. the mutation frequency) and a graph denoted as $g(V, E)$.

HotNet2 detects gene groups rather than single genes strongly connected to cancer progression. Gene groups are identified by detecting "strongly connected components" in the inferred network. A strongly connected component contains fully connected nodes. HotNet2 uses a statistical test to determine sub-networks that are significantly enriched with somatic mutations across multiple cancer types in specific functional categories, such as protein complexes.

MEXCOwalk [169] builds a data-specific PPI network that conserves only mutated genes as nodes. Edges in the network are weighted to reflect the combination of network connectivity, mutual exclusivity, and coverage. Modules of cancer driver genes are identified by refining the *strongly connected components* in the network.

Strongly connected components are defined in the same manner as they are for HotNet2.

Unlike HotNet2, MEXCOwalk allows the initial models to be split and the sub-networks to be reconfigured during the refining process. In this sense, HotNet2-like modules are *seed* modules that are used as starting points for an optimisation process to create new sub-networks. The method hypothesises that refined models are cancer driver modules because of the joint regulation of the genes within them.

Although methods for detecting multi-gene modules provide valuable information on the associations that are strongly related to cancer, such associations are not necessarily linked to the causes of cancer. **DriverGroup** [170] is one recently developed method that addresses this issue by searching for groups of genes that jointly drive cancer progression. This method constructs a data-specific miRNA-TF-mRNA network using a PPI network and additional biological knowledge to determine the topology of the connections in the inferred network. It then applies control theory to identify critical nodes in the network that represent the putative drivers. The method identifies cancer driver gene groups by detecting groups of critical nodes that have a significant influence on proliferation genes.

Most of the current methods used for gene module detection, such as MEMo, iM-CMC, HotNet2, and MEXCOwalk, are prone to finding functionally related genes involved in cancer rather than groups of cancer drivers. This is because not all gene dysregulations in a multi-gene module cause cancer to develop, as mutations in a few members of a module may be sufficient to initiate cancer progression [144].

In this thesis, the current limitations of gene group-based methods are addressed via the development of a custom dynamic causal model that identifies groups of genes responsible for jointly driving cancer. The fact that gene regulation changes dynamically in biological processes because of the influence of several factors (e.g. the cell cycle, metabolic status, and epigenetic changes) is taken into account. Specifically, the gene regulation dynamics that occur *during* cancer are modelled to discover the behaviour of the systems formed by the gene groups that drive cancer.

The following chapters discuss the three primary outcomes of the research performed for this thesis. Chapter 4 presents a dynamic causal inference method constructed to identify the critical interactions characterising biological processes that occur during cancer progression. This approach allows the user to characterise individually dysregulated biological processes as well as the overall cancer. Chapter 5 describes a framework for identifying *dynamic cancer drivers*, which are defined as genes that drive one or more of the altered biological processes responsible for cancer progression. Chapter 6 discusses a causal approach for modelling the dynamics of the interactions and collaborations between key genes that jointly drive cancer progression. Finally, Chapter 7 summarises the outcomes and implications of this work and discusses the future research directions.

Chapter 4

A pseudotemporal causality approach to identifying miRNA-mRNA interactions during biological processes

The discussion in the previous chapters pointed out that most current methods for identifying gene regulatory networks (GRNs) in cancer neglect the dynamics of genes during cancer progression. Consequently, GRNs inferred by those methods are static representations, and thus they may not accurately represent the regulatory relationships that drive the altered biological processes that cause cancer. Furthermore, the vast majority of analyses focus on coding genes and overlook other regulatory relationships that are significant in cancer, such as microRNA (miRNA) - messenger RNA (mRNA) regulatory relationships.

In this chapter, a novel framework called *pseudotemporal causality* (PTC) is presented. PTC identifies regulatory relationships that manifest in biological processes throughout cancer development. Specifically, the method identifies miRNA-mRNA regulatory relationships that characterise relevant biological processes in cancer. The regulatory relationships are characterised by considering the underlying temporal aspect of the disease during the inference process. The experiments showed that this method significantly outperformed existing methods in detecting miRNA-mRNA regulatory relationships.

The content of this chapter is based on the paper entitled "A pseudotemporal causality approach to identifying miRNA–mRNA interactions during biological processes" [86], published in *Bioinformatics*.

4.1 miRNA-mRNA regulatory relationships

Identifying the regulatory relationships between miRNAs and mRNAs that characterise a biological process is a crucial task. For example, miR200c regulates ZEB1 during the epithelial-to-mesenchymal transition (EMT), a key process in cancer metastasis, and therefore miR200c and its interaction with ZEB1 can be used in the treatment of cancer metastasis [171]. Understanding the miRNA-mRNA interactions that characterise different biological processes provides insights into the biology of diseases and aids the design of disease treatments.

A large number of computational methods that aim to find miRNA-mRNA relationships in data have been developed. The first wave of methods examined the negative correlation in the expression levels of miRNAs and mRNAs, as the majority of miRNAs down-regulate their target mRNAs (see [172] for a review).

The second wave of methods focused on the causal relationships between miRNAs and mRNAs, and thus were aimed at excluding the spurious relationships between miRNAs and their target mRNAs [130, 173]. The third and current wave of research is focused on the relationships between miRNAs and mRNAs relative to other types of molecules, (e.g. long non-coding RNAs). An example of this category is the miRNA sponge interaction networks, for which researchers investigate the competition between different RNAs (including mRNAs, lncRNAs, and pseudo-genes) in fighting for a place to bind with miRNAs [174, 175].

Most of the current computational methods for discovering miRNA-mRNA interactions use the data generated for one specific snapshot of a biological process and ignore the dynamics or temporal aspect of biological processes. For example, several studies [176, 72] aimed to find the miRNA-mRNA interactions by analysing gene expression data obtained from cancerous tumours. In these studies, the data employed was collected from different patients at a single point in time. Other works [128, 177, 178] used genetic information (e.g. gene expression) for two different points in time or conditions to contrast the findings between different biological states (e.g. cancer vs normal, epithelial vs mesenchymal).

However, understanding what is happening *during* a biological process (e.g. how genes interact with each other while cells transform from a normal stage to an invasive stage) is ideal. Time series data collected during a biological process can be used to infer the miRNA-mRNA interactions that characterise the process [179, 102], but this dynamic analysis is rarely performed in practice. Given a biological process of interest, gene expression data needs to be collected for both miRNAs and mRNAs at different points in time in the process. Because of the costs of the experiments, there are very few of these datasets available. Therefore, using other types of data and developing novel methods to examine the miRNA-mRNA interactions that occur during a biological process are required.

Recent developments in single-cell sequencing and the concept of *pseudotime* have opened the door for developing novel methods that reveal what is happening during biological processes [79]. Single-cell sequencing techniques quantify gene expression profiles of transcripts in individual cells. In addition, the pseudotime approach utilises the expression (extracted from cross-sectional data) of the markers of a biological process to order the cells along the progression of the biological

process. A wide range of studies has shown that such an order is intrinsically temporal and reflects the progression of a biological process [79, 88, 180, 84].

In this thesis, a novel approach called *pseudotemporal causality (PTC)* is developed. *PTC* elucidates the miRNA-mRNA interactions that occur during biological processes using gene expression data with the expression profiles of matched miRNAs and mRNAs in the same cells or samples. Given a biological process, *PTC* first transforms the matched miRNA and mRNA single-cell gene expression data into pseudotime data using the marker genes of the biological process. *PTC* relies on the causal invariance property [13, 60] to find the causal relationships between miRNAs and mRNAs using the pseudotime data. *PTC* was applied to the single-cell dataset from [77] and the bulk data from [75] (refer to sections 4.3.1 and 4.3.5 for a detailed description). In both cases, VIM, an EMT marker, was used to define the pseudotime.

The results showed that *PTC* significantly outperformed the benchmark methods in identifying experimentally confirmed miRNA-mRNA interactions using either single-cell or bulk data. This suggests that the temporal information on a biological process is useful for revealing the miRNA-mRNA interactions that characterise the biological process.

4.1.1 Invariance property of causal structures for causal inference from sequential data

One of the most prominent properties of causal structures is the fact that when all "direct causes" $\mathbf{X}$ of a target variable of interest Y are considered, the conditional probability of Y given $\mathbf{X}$ will not change even under interventions affecting any variable other than Y [13]. This idea is referred to as the *invariance property* of a causal structure, which is closely linked to causality and has been traditionally referenced in the field as *invariance* [13], *autonomy* [181], *modularity* [182], or *stability* [183].

Theoretically, when considering a target variable Y and a set of covariates $X^1, ..., X^d$, it is possible to infer the complete set of *causal predictors* $X^1, ..., X^d$ (i.e. the causes of Y) by testing whether or not the *invariance property* holds [60]. This idea, known as *invariant causal prediction* (ICP), can be implemented in settings for which data

from Y and its covariates from different experimental settings or regimes are available. The original method proposed by Peters et. al. [13] assumes that data comes from different "environments" (such as different experimental conditions) $e \in \mathcal{E}$, and have an independent and identically distributed sample of (X^e, Y^e) in each environment.

Provided that the full set of causes of Y (usually denoted as S^*) is a subset of the covariates $X^1, ..., X^d$ (i.e. $S^* \subseteq \mathbf{X}$), the ICP method guarantees (in terms of statistical confidence) that it is possible to estimate a set of variables $S \subseteq S^*$ such that S is a set of causal predictors (with a controllable large coverage probability of $1 - \alpha$) [13, 60]. The essential assumption of ICP can be formally summarised as requiring that the conditionals $Y^e|\mathbf{X}^e_{S^*}$ and $Y^e|\mathbf{X}^f_{S^*}$ are identical for all environments $e, f \in \mathcal{E}$ [13]:

$$Y^e|\mathbf{X}^e_{S^*} \equiv Y^e|\mathbf{X}^f_{S^*} \qquad \forall\, e, f \in \mathcal{E}. \tag{4.1}$$

Although the ICP idea is useful, it is restrictive because it considers only the static version of the data. Because of the statistical nature of the test in ICP, it is possible (and expected) that more than one combination of predictors will not violate the invariance property. To achieve the statistical guarantee, the approach needs to compensate for the limitations of using static data. This translates into the (extremely) conservative strategy of keeping only the intersection set of all combinations of predictors that do not violate the property of invariance. This strategy often results in the method inferring the empty set or a small subset of the true set of causes.

Including the temporal order characteristic of the time series allows the invariant structures (i.e. conditionals) that do not change over time to be searched. This not only has the advantage of allowing the invariance property to be used in scenarios without knowledge of different environments, but also allows the utilisation of temporal precedence to improve the validity of causal inference. Pfister et al. proposed the *sequential ICP* approach [60], which reformulates the *invariance property* test in terms of the structural invariance of a regression model.

This reformulation requires n sequential data points from the target **Y** and d predictors that conform the matrix **X**; this is denoted by **(Y,X)** $= (Y_t, X_t)_{t\in\{1...n\}} \in \mathcal{R}^{n\times(d+1)}$. Using this kind of sequential data, it is possible to define an *invariant set* with respect to **(Y,X)** as a set $S \subseteq \{1, ..., d\}$ that fulfills the following:

- There exist parameters $\mu \in \mathcal{R}, \beta \in (\mathcal{R}\backslash\{0\})^{|s|\times 1}$ and $\sigma \in \mathcal{R}_{\geq 0}$ such that $\forall t \in \{1, ..., n\}: \quad Y_t = \mu + X_t^S\beta + \varepsilon_t$ and $\varepsilon_t \perp\!\!\!\perp X_t^s$

- $\varepsilon_1, ..., \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Similar to the static version of ICP, the outcome of the *sequential ICP* method corresponds to an intersection. Specifically, in the version proposed in [60], the outcome corresponds to the intersection of all sets that fulfills the aforementioned definition of *invariant set*.

The direct implementation of the sequential version of ICP was, during the course of this thesis, found to be impractical in the field of cancer biology for two main reasons. First, there are very few time series in cancer samples with sufficient information to allow a reliable inference. Second, although the version developed by Pfister et al. lifts the restrictions regarding explicit knowledge of different environments, the method is still very conservative. Thus, when it is applied to real data, the method frequently returns the empty set as the solution subset.

In this thesis, the idea of searching for structures that are *invariant* over a period of time formed the basis the development of a framework that utilises causal inference to unveil causal relationships during cancer progression. Specifically, this method aims to reveal causal relationships between microRNAs and mRNAs by using the *invariance property* of causal structures that drive core biological processes in cancer. Such relationships are a critical component of the GRNs involved in carcinogenesis and cancer development. In particular, using the *invariance property* in the context of sequential data in cancer was found to significantly enhance the detection of regulatory relationships that shape the biological processes that occur in cancer.

4.2 Methods

4.2.1 Problem definition and method outline

Based on the premise that a biological process can be characterised by miRNA-mRNA interactions, it is necessary to develop a method capable of inferring those interactions from data obtained while the biological process is occurring. Ideally, the dataset would consist of time-series data obtained from experiments carried out during the biological process of interest. However, collecting time-series experimental data is costly and sometimes infeasible. As a result, most of the currently available experimental data are static data recorded at a few points in time during a biological process.

PTC is based on the assumption that a biological process, such as cancer progression, is governed by an underlying causal mechanism. This presumption defines the two main research problems to be solved by this method. First, temporal data corresponding to different points in time during the process are required to infer the miRNA-mRNA relationships that characterise the process. Second, because miRNA-mRNA relationships are considered causal, a technique for testing causality during the biological process is also required.

PTC solves the temporal data requirement by performing a pseudotime analysis and transforming static data (in this case, miRNA/mRNA gene expression) into their pseudotime version. To determine the causal relationships between miRNAs and mRNAs, this method presumes that the relationship between an mRNA (the target) and a set of miRNAs (the regulators) can be modelled as a linear system in which the effect or response variable corresponds to the gene expression of the mRNA and the cause or predictor variables correspond to the expression set of miRNAs.

As shown in Figure 4.1, the overall procedure of the proposed *PTC* method consists of three phases:

- **Phase 1: Pseudotime analysis.** *PTC* takes a (static) gene expression dataset containing samples of matched miRNA and mRNA gene expressions and converts it into a time series dataset. The conversion is performed via a

pseudotime analysis to reorder the samples according to the pseudotime established using a well-recognised EMT marker, VIM (details are in section 4.2.2).

- **Phase 2: Gene selection**. This phase consists of two steps: a pre-selection process and a process to define the set of plausible predictors or causes for an mRNA. In the first step, the top-ranked miRNAs and mRNAs are selected based on the median absolute deviations (*MADs*) of their gene expressions.

  In the second step, for each of the selected mRNAs, a further analysis is carried out to determine the miRNAs that can biologically interact with the mRNA. Only the miRNAs that can target the mRNA as predicted by TargetScan [133] are kept as the plausible predictors or causes of the mRNA.

- **Phase 3: Identification of miRNA-mRNA**. In this phase, *PTC* employs the *decoupled test* [60] to determine whether or not the set of plausible predictor miRNAs of an mRNA obtained in Phase 2 are the real causes or regulators of the mRNA. To determine the causal relationships, *PTC* tests if the relationships violate the "causal invariance" property during the biological process by using the time series data obtained above with the gene selection in Phase 2 (details are in section 4.2.4).

4.2.2 Pseudotime analysis

Because most gene expression datasets contain only a snapshot of the gene expression, it is necessary to transform the data for temporal analysis. The proposed method is based on the assumption that dynamic causation can be inferred from the underlying dynamic in the temporal data. In general, the aim is to find the mRNAs that progressively undergo gene expression changes over time, and to use these genes as pseudotemporal units to model the different trajectories of all genes in the genomic data during cancer progression.

In the experiments carried out for this thesis, the EMT process was selected as the biological process. VIM was used to define the pseudotime because VIM is a well-known EMT marker [184, 185]. During the EMT process, the gene expression

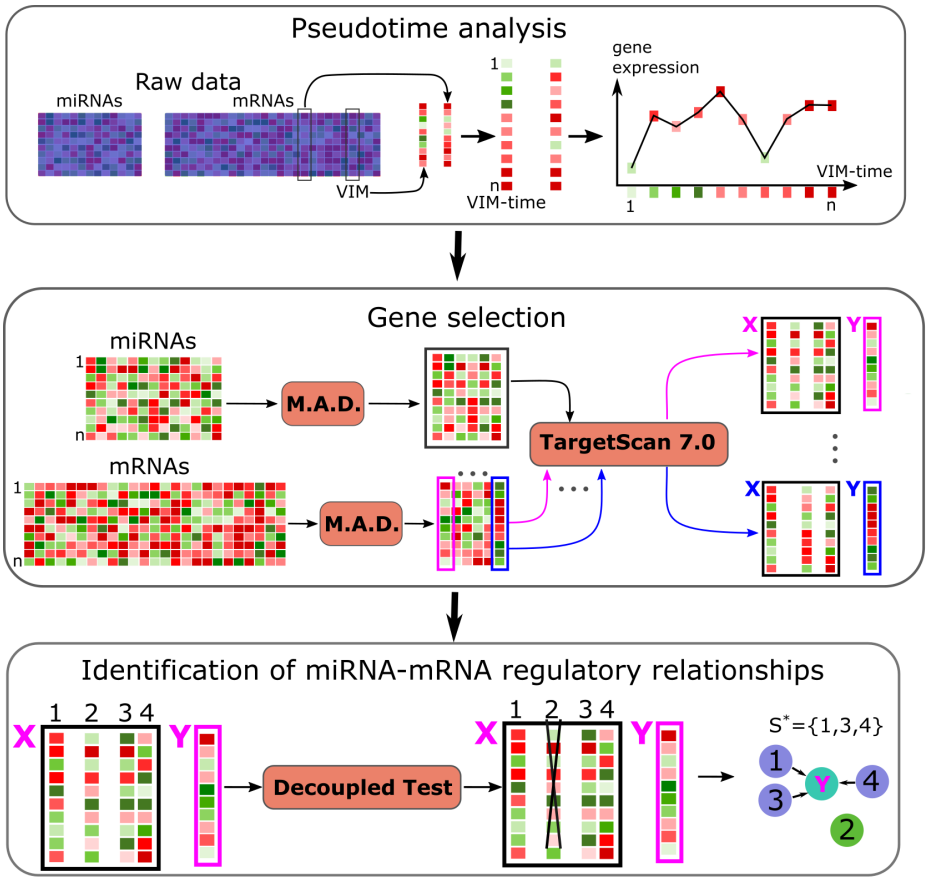


Figure 4.1: **Summary of the proposed method, *pseudotemporal causality (PTC)*.** PTC is comprised of three phases: (1) Pseudotime analysis. miRNA-mRNA static gene expression is processed to create time series data that follows a pseudotime order. In the experiments, VIM was selected as the biomarker to create the pseudotime. (2) Gene selection. The MAD and TargetScan 7.0 (predictions of conserved sites) are employed to determine a set of miRNAs as plausible predictors of each of the **k** genes. (3) Identification of miRNA-mRNA regulatory relationships. Violations to the invariance property are assessed to find the causal parents (predictors) of each gene.

level of VIM increases [186]. This implies that the ascending ordering of the expression of VIM in an EMT dataset provides a pseudotime for EMT (called "VIM-Time" for the rest of this chapter). Hence, to convert the static matched mRNA and miRNA samples into temporal data, in the proposed method the matched mRNA and miRNA samples are sorted based on the VIM-Time (i.e. the input data matrix is sorted in ascending order according to the column containing VIM's expression values). As a result, changes in gene expression for all genes were described as a progression in terms of VIM-Time. This data transformation incorporates a temporal dimension into the data, which *PTC* uses to identify causal relationships during the EMT process.

4.2.3 Gene selection

A comprehensive analysis that determines all parents of a response variable by testing the invariance property could require evaluating all possible combinations of the plausible predictors. Peters et al. proposed a method for determining the set of true causes of a variable as the intersection of all sets that do not violate the invariance property [13]. Finding the intersection set could require testing an enormous number of combinations. This can easily be infeasible if the number of plausible predictors is large. For example, if an mRNA has 30 plausible predictors, in the worst case of Phase 3 for this single mRNA, over a billion tests would need to be conducted because there are 2^{30} possible combinations for 30 miRNAs.

Considering this fact, in the proposed method the gene selection phase includes a heuristic procedure to reduce the number of predictors for each gene. During the first step of this phase, a subset of miRNAs and a subset of mRNAs are selected. This selection is based on the hypothesis that miRNAs and mRNAs with the most expression variability are more likely to be relevant to regulation processes in cancer. For *PTC*, the MAD is used as the measure for gene expression variability. This selection was performed using the function *"FSbyMAD"* from the *"CancerSubtypes"* R package [187]. In the experiments, the parameters for *"FSbyMAD"* were: cut.type = "topk", value = 30 or 1500 when selecting miRNA or mRNA, respectively. These choices replicated the selection process described in [75].

After this step, *PTC* employs a heuristic to further reduce the set of plausible predictors for each mRNA to the set of miRNAs that can bind to that gene as predicted by TargetScan 7.0. The set of plausible predictors for each gene is restricted to a tractable size for the testing conducted in Phase 3.

4.2.4 Identification of miRNA-mRNA regulatory relationships

Causal relationships have an *invariance property*. Specifically, for a causal structure, the conditional distribution of a variable (called a target or effect variable in this case), given all its direct causes, remains the same under interventions on any other variables in the structure except the response variable.

The property of *invariance* can be exploited to determine whether a set of plausible predictors is the set of direct causes of a response variable. In general terms, given

a set of plausible predictors and one response variable, violations to the invariance property can be assessed using data obtained from different environments (which may be associated with interventions on different variables, as described above). If a set of plausible predictors violates the invariance property across different environments, it can be inferred that the set of plausible predictors contains variables that are not direct causes of the response variable.

4.2.4.1 Testing violation of the invariance property

Phase 3 of the proposed method determines the existence of a causal link between a set of miRNAs (parents) and an mRNA by assessing violations of the invariance, adapting the definition of "invariant set" given in [60] as follows.

Definition 4.2.1. (Invariant set). Given a column vector $\mathbf{Y} := ((Y_t)^T)_{t\in\{1,..,n\}}$ containing the gene expression for n temporally ordered samples of the mRNA considered as the response variable of interest, and a matrix $\mathbf{X} := ((\mathbf{X}_t^1)^T, ..., (\mathbf{X}_t^d)^T)_{t\in\{1,...,n\}}$ containing the gene expression for n temporally ordered samples of the d miRNAs considered as plausible predictors of the mRNA (Phase 2 of *PTC*). Furthermore, for any set $S \subseteq \{1, ..., d\}$, the matrix $\mathbf{X}^S$ contains only the columns of $\mathbf{X}$ indexed by S. S is an "invariant set" with respect to $(\mathbf{Y}, \mathbf{X})$ if there are parameters $\mu \in \mathbb{R}$, $\boldsymbol{\beta} \in (\mathbb{R} \setminus \{0\})^{|S|\times 1}$, and $\sigma \in \mathbb{R}_{>0}$ such that

$$\begin{aligned} \forall t \in \{1, .., n\} : \quad & Y_t = \mu + X_t^S \boldsymbol{\beta} + \varepsilon_t \\ & \varepsilon_t \perp\!\!\!\perp X_t^S \text{ and } \varepsilon_1, ..., \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \end{aligned} \tag{4.2}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, μ is the regression-independent term, and $\varepsilon_1, ..., \varepsilon_n$ are independent and identically distributed random noise functions, which follow a normal distribution with a mean of zero and standard deviation σ.

Based on the above definition of an invariant set, if S does not violate the invariance property, $\forall t \in \{1, .., n\}$, μ, $\boldsymbol{\beta}$, and σ^2 remain the same. This feature allows the use of m adjacent time points with $m < n$ to create segments of time series and use them as environments, instead of using data from different experimental settings. The segments of the time series that simulate a single environment must have the same number of time data points, and they must be synchronized.

PTC searches for the following violations in the models of Y (as in Eqn. 4.2) inferred from different segments of the time series: (i) significant differences between the vectors of regression coefficients ($\boldsymbol{\beta}$) and/or (ii) differences in the noise variance (σ^2). For the experiments described in section 4.3, the time series was divided into two similar long segments to assess the differences in the invariant set models.

The *decoupled test* [60], so named because it tests the two violations described above independently, is employed to verify if a set of miRNAs violates the invariance described above. To conduct the test, the function *seqICP.s* from the seqICP R package [60] was used with a significance level of $\alpha = 0.02$. The output of this test produced the lowest of the p-values obtained after testing the two violations described above. In the experiments, a p-value of < 0.02 implied at least one violation to the invariance property.

4.2.4.2 Identifying miRNAs regulating a mRNA

For each mRNA of interest (Y) and its set of plausible predictor miRNAs obtained during Phase 2, *PTC* conducts the invariance property test described above on each subset of the set of plausible predictors to find all the subsets of miRNAs that do not violate the invariance property. Then *PTC* outputs the union of all the subsets that possess the invariance property (denoted as S^*) as the set of direct causes or regulators of Y. The decision to use the union of all subsets of plausible predictors that do not violate the invariance property is based on the fact that *PTC* only admits as plausible predictors the miRNAs that can biologically target the gene.

To make *PTC* more efficient, the invariance property test starts with testing the largest subset of the d plausible predictors (i.e. $\{1, .., d\}$). Then the subsets of size (d-1), (d-2), and so on are tested, and finally the subsets of size 1 are tested. If $S' \subseteq \{1, 2, ..., d\}$ is tested to satisfy the invariance property, then all subsets of S' will not be tested, as *PTC* outputs the union of all subsets of $\{1, 2, ..., d\}$ satisfying the invariance property. For example, assume that a gene has three plausible predictors, {1, 2, 3}. In this case, the tests would be conducted in the following order: {1,2,3} {1,2} {1,3} {2,3} {1} {2} {3}. If the {1,3} test does not violate the invariance property, then all of its subsets (i.e. {1} and {3}) will not be tested.

The final output of *PTC* contains all the identified miRNA-mRNA pairs and their ranking scores. The TargetScan 7.0 Context++ scores was used for all conserved

miRNA sites, and the score of each miRNA-mRNA pair was calculated as the average of all context scores of that pair across all conserved sites. Because context scores are negative numbers, the more negative the average is, the higher the rank the miRNA-mRNA pair has in the proposed method.

The performance of *PTC* was measured by the percentage of miRNA-mRNA relationships inferred by *PTC* that have been confirmed experimentally. For this evaluation, a collection of confirmed interactions was assembled by combining information from several databases, including miR-Tarbase 6.1 [188], Tarbase 7.0 [189], and miRWalk 2.0 [190]. An inferred interaction is considered experimentally confirmed if it can be found in at least one of these databases. The performance of *PTC* was also compared to that of traditional miRNA-mRNA interaction inference methods (including *Pearson* [191] and *Lasso* [192]) using the same performance indicator (i.e. proportion of predicted miRNA-mRNA relationships confirmed experimentally). "GO biological processes" and "KEGG pathways" analyses were performed on the miRNA-mRNA causal relationships determined by *PTC* to assess their relevance to cancer and EMT.

4.3 Results and discussion

4.3.1 Identified miRNA-mRNA interactions in EMT with single-cell data

Single-cell techniques normally destroy the cell when measuring the gene expression. Thus, it is difficult to measure the miRNA and mRNA expressions simultaneously. As a result, the technique of quantifying the expression of miRNAs and mRNAs in single cells at the same time is still in its infancy. Only one study [77] has provided matched miRNA and mRNA expressions for single cells, but with a very limited number of cells (20 cells). In this dataset, the gene expression corresponds to single human acute myeloid leukemia cells.

PTC was applied to this single-cell dataset [77] to find miRNA-mRNA interactions in EMT. This dataset provides a total of 20 cells that were each split in half in order to obtain paired miRNA-mRNA gene expression data (miRNA expression from one half of the cell and mRNA expression from the other half of the same cell). Cell number 6 was excluded from the analysis because it does not have paired miRNA-mRNA gene expression data. Therefore, the 19 remaining cells, each with paired

entries, were employed. After the data was processed (please see Appendix A, section A.1), the output dataset contained 2822 miRNAs and 23141 mRNAs from 19 single cells.

According to the processed dataset, *PTC* predicted 569 miRNA-mRNA interactions, as visualised in Figure 4.2. Experimentally verified interactions are shown in red. All other inferred regulatory relationships are represented in blue, and their strength is indicated by the colour intensity (the darker the line, the stronger the predicted interaction).

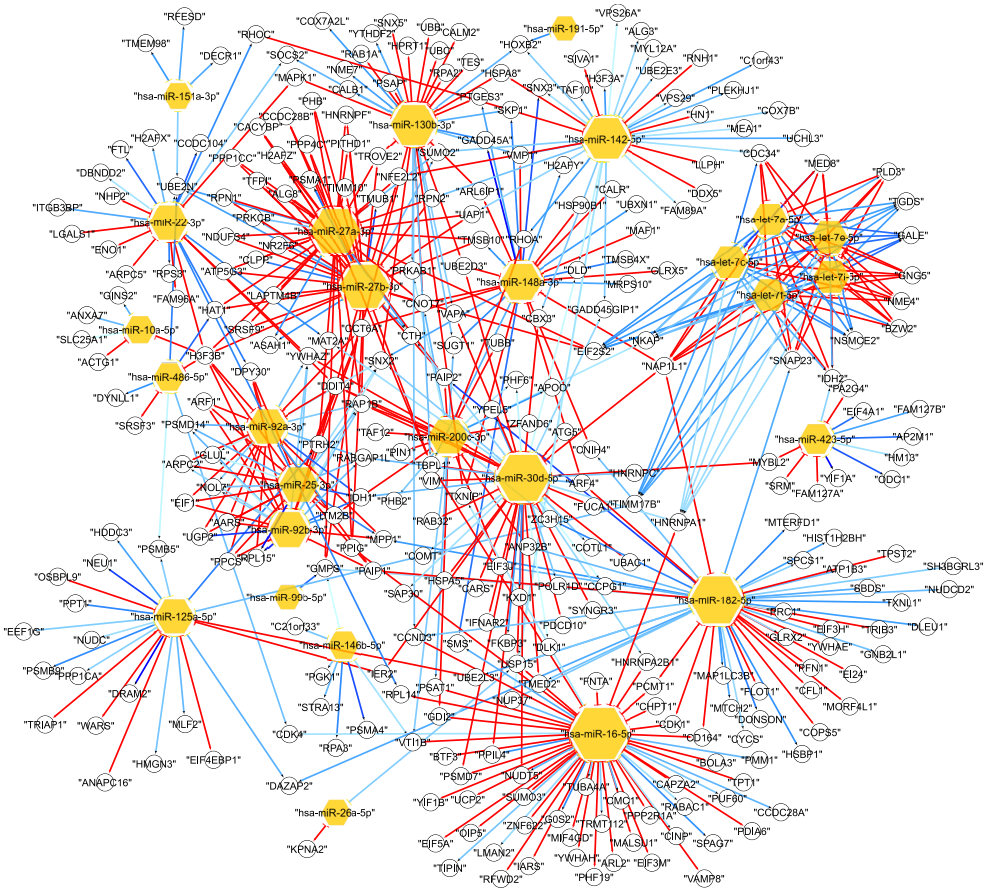


Figure 4.2: **miRNA-mRNA regulatory relationships discovered by *PTC*.** Experimentally confirmed interactions are shown in red. All other predicted interactions are shown in blue; the strengths of these causal links are represented by the colour intensity (the darker the line, the stronger the link). A larger miRNA node indicates that the miRNA has more predicted targets. The most influential nodes as determined by PTC were ”hsa-miR-16-5p”, ”hsa-miR-182-5p”, “hsa-miR-30d-5p”, “hsa-miR-27a-3”, and “hsa-miR-27b-3p”, with 55, 47, 46, 40, and 40 inferred regulatory relationships, respectively. Several studies have linked these miRNAs to different types of cancer, such as leukemia [193] and colorectal cancer [194].

The number of genes predicted to be regulated by an miRNA is indicated by the size of the miRNA node (i.e. a larger miRNA node has more predicted targets). The most influential nodes identified were "hsa-miR-16-5p", "hsa-miR-182-5p", "hsa-miR-30d-5p", "hsa-miR-27a-3p", and "hsa-miR-27b-3p", with 55, 47, 46, 40, and 40 inferred regulatory relationships, respectively. Several studies have linked these miRNAs to different types of cancer. The "hsa-miR-16-5p" node appears to play a significant role in chronic lymphocytic leukemia [193] and periocular sebaceous gland carcinoma [195]. The "hsa-miR-182-5p" node has been linked to colorectal cancer [194], drug resistance in breast cancer cell lines [196], epithelial ovarian cancer [197], and lung adenocarcinoma [198]. The "hsa-miR-30d-5p" node has been linked to non-muscle-invasive bladder cancer [199], while the "hsa-miR-27a-3p" node has been associated with progression of spinal cord glioma and intrahepatic cholangiocellular carcinoma [200]. Finally, the "hsa-miR-27b-3p" node has been related to tumour suppression in lung cancer [201] and gastric cancer [202].

4.3.2 Experimental confirmations of interactions discovered by *PTC*

Among the 569 miRNA-mRNA regulatory relationships discovered by *PTC*, 294 pairs are experimentally confirmed interactions. For each group of the top interactions predicted (following the ranking explained in section 4.2.4.2), the ratio of confirmed interactions increases as follows (see Table 4.1): 121 out of 200 (60.5%), 93 out of 150 (62%), 65 out of 100 (65%), and 37 out of 50 (74%).

Table 4.1: **Summary of performance results for *PTC*, Pearson, and Lasso.** Comparison of confirmed interactions in the top **k** lists inferred by each method.

	PTC		Pearson		Lasso	
Top	Confirmed	Percentage	Confirmed	Percentage	Confirmed	Percentage
200	121	60.5%	25	12.5%	28	14%
150	93	62%	21	14%	21	14%
100	65	65%	16	16%	14	14%
50	37	74%	10	20%	8	16%

For comparison purposes, this dataset was analysed using the *Pearson* and *Lasso* methods from the miRLAB R package [203]. These two methods are commonly used for inferring regulatory relationships. The inferred interactions found by these two methods were verified using the databases of confirmed interactions previously mentioned. The performance of *PTC* was significantly better than the *Pearson* and *Lasso* methods (as indicated by Table 4.1).

4.3.3 Identification of interactions relevant to EMT using *PTC*

Because *PTC* is based on the idea that there are miRNA-mRNA regulatory relationships that characterise a biological process (the EMT process was used in the experiments), the relevant interactions during EMT that were inferred by our method were sought. To determine if a miRNA-mRNA interaction discovered by *PTC* was relevant to EMT, a list of EMT miRNAs/mRNAs that have been experimentally confirmed as relevant to the EMT process was assembled (refer to Table S1A in [204] and Table S1 in [164]). In the experiments, a miRNA/mRNA was considered relevant to EMT if it could be found in the aforementioned list.

The numbers of EMT miRNAs in the top 200, 150, 100, and 50 miRNA-mRNA interactions discovered by *PTC* were 12, 11, 10 and 9, respectively. A significant number of the miRNA-mRNA interactions discovered by *PTC* involved these EMT miRNAs. The percentage of relationships inferred by *PTC* that involved those EMT miRNA biomarkers across the tops lists were: 44.5% (89 out of 200), 44% (66 out of 150), 46% (46 out of 100), and 44% (22 out of 50). The inferred pairs containing EMT miRNAs were also highly ranked by the proposed method. It is logical that the number of EMT miRNAs in pairs inferred by the proposed method was high and persistent because the proposed method employed the temporal information on the EMT process to determine the causal regulatory relationships that occur during the EMT process in cancer.

PTC was able to discover EMT-related miRNA-mRNA interactions that other methods fail to discover. For example, the experimentally confirmed interaction (hsa-miR-16-5p - VAMP8) was discovered by *PTC*. The classical correlation-based *Pearson* and *Lasso* methods failed to detect this relationship because of its low correlation and "non-significant" p-value. This suggests that *PTC* is able to detect causal relationships that correlation methods are unable to identify because it uses temporal information. Figure 4.3 shows the gene expression of the pair (hsa-miR-16-5p - VAMP8) as a function of the VIM-Time. An apparent correlation can be observed in certain areas of the plot (e.g. the VIM-Times between 5 and 10), and some areas indicate negligible correlation (e.g. the VIM-Times between 10 and 15).

"GO biological processes" and "KEGG pathways" analyses were also conducted for the miRNAs and genes in the interactions discovered by *PTC* to examine their

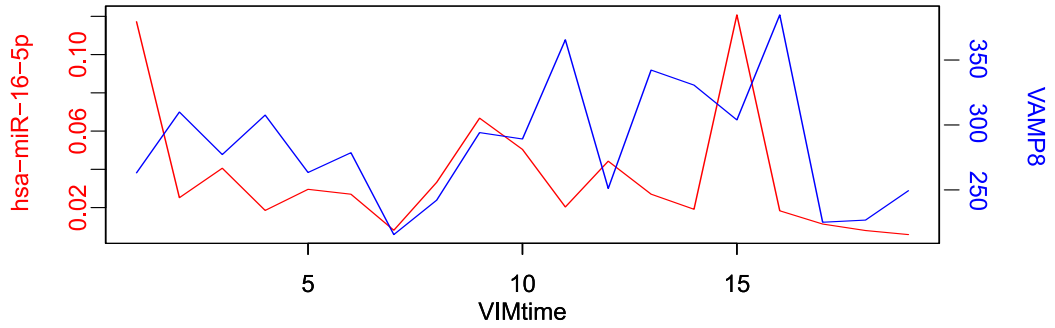


Figure 4.3: **Gene expression of the pair (hsa-miR-16-5p - VAMP8) as a function of the VIM-Time.** The red line shows the gene expression of "hsa-miR-16-5p" and the blue line shows the gene expression of VAMP8. This pair exhibited a Pearson's product-moment correlation of 0.04591796 and a p-value of 0.8519262; hence, this regulatory pair is rejected by the Pearson method. However, this regulatory relationship has been experimentally confirmed, and it is accepted by the proposed method. Therefore, *PTC* can infer miRNA-mRNA interactions relevant to EMT that correlation methods fail to discover.

relevance to cancer and EMT. These analyses were performed by using the *Enrichr* web tool [205]. A large number of mRNAs present in the regulatory pairs identified by the proposed method are associated with KEGG pathways and GO biological processes relevant to the EMT process and cancer (refer to Appendix A and Table A.1).

4.3.4 Stability of *PTC* at different pseudotimes

To test the stability of *PTC* with respect to the pseudotime transformations, the *Wanderlust* algorithm [79] was used to identify EMT pseudotime. *Wanderlust* outcome is a pseudotime based on pre-defined EMT markers from 315 general EMT markers in cancers [204]. In the single-cell mRNA data employed in this experiment, there were 145 of those EMT markers. Thus, pseudotime was estimated by running *Wanderlust* on these 145 EMT markers.

There are four key parameters in *Wanderlust*: (1) the number of nearest neighbours k, (2) the shared nearest neighbour snn, (3) the number of neighbours selected for each node in a k-nearest neighbours graph l, (4) and the number of l-out-of-k-nearest neighbours graphs ng. In the experiments, the parameter k was set to 4 to maintain a reasonable proportion between the number of cells (19) and the number of neighbours per node. Parameters l and ng must be positive integers lower than k. They both were set to 2 because it was the median of the possible options (i.e.

$\{1, 2, 3\}$). Because all other parameters were fixed, the parameter *snn* was required to be 0.

The performance of *PTC* when *Wanderlust* pseudotime was employed was very similar to the performance of *PTC* when VIM-Time was used. Using *Wanderlust*, 570 miRNA-mRNA relationships were identified (only one more than the 569 discovered when using VIM-Time). The confirmed interactions among the top miRNA-mRNA interactions discovered were also similar: 122 out of 200 (vs 121 for VIM-Time), 94 out of 150 (vs 93 for VIM-Time), 65 out of 100 (vs 65 for VIM-Time), and 37 out of 50 (vs 37 for VIM-Time).

Only one experimentally confirmed pair inferred by *PTC* when VIM-Time was used, (hsa-miR-10a-5p - H3F3B), could not be detected when *Wanderlust* was employed. Conversely, only two pairs detected by *PTC* when *Wanderlust* was used, (hsa-miR-27a-3p - ID3) and (hsa-miR-27b-3p - ID3), were not detected when VIM-Time was employed. These results suggest that *PTC* is robust and stable against different EMT pseudotime data transformations.

4.3.5 Performance of *PTC* relative to benchmark methods for bulk gene expression data

Pseudotime is also useful for recovering certain aspects of the latent dynamics of a process from bulk datasets [84]. For this reason, and because most of the current methods for inferring gene regulatory relationships use bulk data, a benchmark test of *PTC*'s inference capabilities was carried out by using bulk data. The test was executed using the bulk miRNA and mRNA expression data employed in [75]. This dataset contains 518 miRNAs and 17403 mRNAs from 503 tumour samples, downloaded from *"The Cancer Genome Atlas Research Network"* (TCGA-BRCA) project (https://www.cancer.gov/tcga). The methods selected for comparison were *Pearson* [191], *Lasso* [192], *idaFast* [130], *jointIDA_direct* [206], and the methods based on *Invariant Causal Prediction (ICP)* (*hiddenICP, hiddenICP pam50, Borda hiddenICP, Borda hiddenICP Pam50*) proposed in [75]. The first four methods were selected because they are widely used in gene regulatory inference. The *ICP*-based methods were selected because they also use the causal invariance property, but with static data.

Table 4.2: **Summary of bulk data performance of *PTC*** compared to the two best non-causal methods, the best (average) method based on static ICP, and PTC. The performance was measured as the number and percentage of experimentally confirmed interactions in each top list. The number of confirmed EMT miRNA and mRNA markers inferred by each method is given in columns "miR_EMT" and "mR_EMT", respectively.

	Pearson				Lasso			
Top	Confirmed	Percentage	miR_EMT	mR_EMT	Confirmed	Percentage	miR_EMT	mR_EMT
200	5	2.5%	4	8	5	2.5%	6	4
150	5	3.333%	4	6	5	3.333%	5	3
100	4	4%	4	4	4	4%	2	2
50	0	0%	3	1	0	0%	1	1

	Hidden ICP pam50				PTC			
Top	Confirmed	Percentage	miR_EMT	mR_EMT	Confirmed	Percentage	miR_EMT	mR_EMT
200	1	0.5%	5	4	25	12.5%	10	2
150	1	0.667%	3	4	22	14.667%	10	1
100	1	1%	3	3	13	13%	9	1
50	1	2%	3	1	9	18%	8	0

Note: The Pearson and Lasso methods are the two not based on ICP with the best performance across all tops. The *hiddenICP pam50* version was selected as the method with the best (average) performance because of the fact it had confirmed predictions in all tops.

In this bulk dataset, a total of 293 miRNA-mRNA regulatory relationships were identified by the proposed method. For all methods, the top 200, 150, 100, and 50 inferred regulatory relationships were selected for analysis. For all the methods but *PTC*, the procedure in explained in [75] provides a rank for each interaction inferred by each method, and it was used in the analysis to determine the top lists for each of the eight methods compared. *PTC* used its own ranking system based on the average of the TargetScan 7.0 Context++ scores for all conserved miRNA sites of inferred miRNA-mRNA pairs across all conserved sites (section 4.2.4.2). Figure 4.4 shows the number of experimentally confirmed miRNA-mRNA interactions inferred by each method (verified using the collection described in section 4.2.4.2).

To provide the time series data required by *PTC*, the expression values of VIM in the bulk dataset were ordered and used as a reference (VIM-Time) to sort the expression values of all other genes; in this way the dataset used in [75] was approximated as a pseudotime dataset. Even with the bulk data, which is a collection of static data from a large number of patients rather than sequential data, *PTC* outperformed the other methods in all top lists (top 200, 150, 100, and 50). As shown

in Table 4.2, *PTC* was able to detect 25 confirmed interactions among the top 200 ($\approx$ 12.5%). The *Pearson* and *Lasso* methods were the runners up, with five confirmed interactions (2.5%).

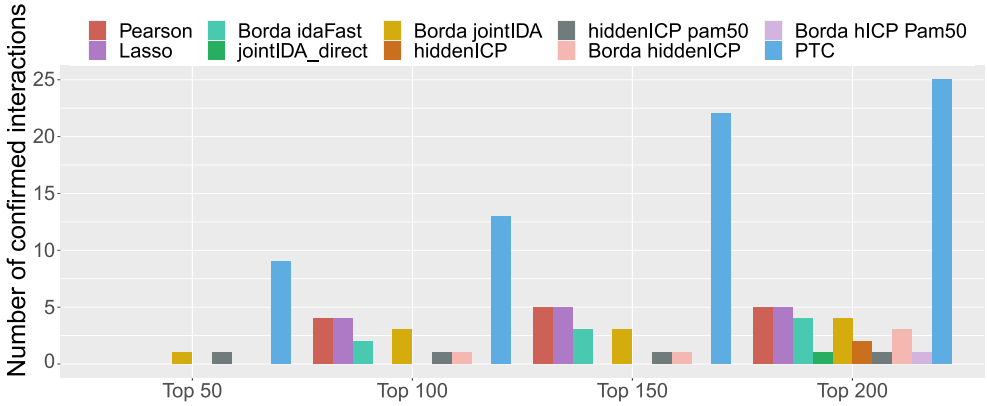


Figure 4.4: **Confirmed interactions found by PTC in bulk data**. The number of interactions found by PTC (shown in blue) is compared to the confirmed interactions found by the other methods analysed (shown in other colours).

Based on the hypothesis that the ranking of the inferred miRNA-mRNA interactions found by each method represents a measure of the strength of the correlation/causal effect, the top 200, 150, 100, and 50 interactions discovered by each method were analysed. An increase in the percentage of confirmed interactions was expected. In general, the percentage of confirmed regulatory relationships increased when fewer miRNA-mRNA pairs were considered in the top list. The proposed method was able to infer confirmed miRNA-mRNA relationships in all the top lists with a rate of at least 12.5%.

The *Pearson* and *Lasso* methods exhibited the second-best performance in the top 200, 150, and 100, but none of the top 50 interactions discovered by the two methods were experimentally confirmed (based on the databases used). The static version of *hiddenICP pam50* was able to determine at least one confirmed relationship in all the top lists. In any case, none of the other analysed methods were able to achieve a percentage of confirmed interactions higher than 4%.

As an additional measurement of performance, the miRNAs and mRNAs found in the inferred regulatory relationships were analysed to verify which of them were confirmed EMT markers. The *Pearson*, *Lasso*, and *PTC* methods were able to detect regulatory relationships involving 10 miRNAs that are confirmed as EMT

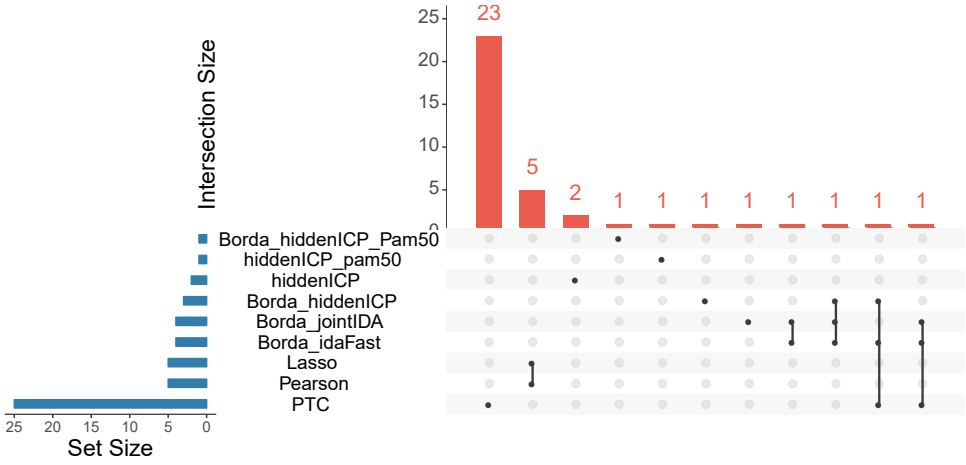


Figure 4.5: **Intersection of the confirmed interactions predicted to be in the top 200 by each method when using bulk data** [75]. The vertical axis represents the size of the intersection. Points below the horizontal axis indicate that the corresponding bar belongs to the method shown on the left. Multiple points under a bar represent the intersected methods. For example, the 25 confirmed pairs found by *PTC* are represented as one point under the bar that is 23 units high and two points (intersections with other methods) under the two bars that are 1 unit high. The visualization was made using *Intervene* [207]

.

miRNAs. Unlike the other methods, *PTC* persistently identified relationships containing the EMT miRNAs, assigning them a high score. As a result, relationships containing these 10 EMT miRNAs were found in the top 200 and 150 lists of *PTC*, as shown in Table 4.2. The intersections between the confirmed interactions inferred by all methods were also investigated. As shown in Figure 4.5, *PTC* was able to infer two regulatory relationships that were also inferred by at least one of the other methods. A total of 23 confirmed interactions were detected exclusively by the proposed method.

4.4 Conclusion

In this chapter, a new approach for discovering causal miRNA-mRNA relationships that characterise biological processes, called *PTC*, was presented. The method assumes that if a set of miRNAs are regulatory agents for the genes that drive a biological process, then the causal relationships between such miRNAs and mRNAs are invariant throughout the process. Based on this assumption, *PTC* performs a data transformation via a pseudotime analysis to create temporally ordered data. In this way, *PTC* identifies the invariance of the causal relationships that exist during the biological process of interest.

In *PTC*, the pseudotime analysis phase warrants correctly and efficiently assessing any violations to the invariance property, as the test is performed on temporally ordered gene expression data. Additionally, the miRNA-mRNA pairs that are not able to biologically interact cannot have a causal relationship. Thus, rejecting these pairs aids *PTC* in determining whether an invariant miRNA-mRNA relationship corresponds to a causal link throughout the process.

Experiments with the data used in this thesis demonstrated the superior performance of the proposed method. *PTC* outperformed all the other methods analysed in terms of the number of confirmed regulatory interactions that were predicted. The results of the tests suggest that using temporal information on biological processes provides additional information that significantly improves the inference of regulatory relationships. The increase in the number of confirmed relationships that were inferred by PTC suggests that the combination of TargetScan 7.0-based heuristics and the invariance property is a powerful and effective tool for uncovering causal relationships that characterise a biological process.

Chapter 5

Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories

In the previous chapter, *PTC*, a pseudotemporal causal method that detects miRNA-mRNA gene regulatory networks characterising core biological processes in cancer, was discussed. In this chapter, a novel causal approach for identifying dysregulated genes that drive such core biological processes is introduced. Dysregulated genes are usually referred to as *cancer drivers* because they cause cancer genesis and they also drive cancer progression. To improve the efficacy of discovering cancer drivers, the concept of cancer driver is expanded to ***dynamic cancer driver***, which is defined as a gene driving one or more biopathological transitions during cancer progression.

The proposed method refers to the fact that cancer should not be considered a single process but a collection of altered biological processes that cause the disease to develop over time. Different drivers of cancer can potentially be discovered by analysing different biopathological pathways. This critical fact is utilised to develop a causal inference method that discovers genes driving one or more core processes during cancer development (i.e. ***dynamic cancer drivers***). The concept of pseudotime is used for inferring the latent progression of samples throughout a biological transition during cancer development and for identifying a critical event when such a process significantly deviates from normal to carcinogenic. Driver genes are inferred by assessing the causal effect they have on the process after this critical event occurs.

The content of this chapter is based on the paper entitled "Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories" [87], published in *Briefings in Functional Genomics*.

5.1 Introduction

Cancer development is traditionally linked to genome-level changes (commonly known as "driver" mutations) that confer proliferative advantage to certain cells, usually leading to a neoplasm [208]. Among the mutations that can be found in cancer tumours, a minimal number promote tumorigenesis, especially in comparison with the vast number of passenger mutations that do not confer superiority to cancer cells [209]. For this reason, most efforts in recent decades have been focused on distinguishing between genes harbouring driver mutations and genes with mutations that do not provide selective advantages to the cell [210, 14].

Because genes interact in complex biological networks, the impact of a genetic aberration can be reflected in other genes besides the mutated gene [211, 212]. As a result, a mutated gene can affect gene products other than its own, even in genes carrying no defects [152]. Taking this into consideration, some methods for discovering cancer drivers have incorporated biological network information into their algorithms [75, 213]. These approaches have become popular in recent years owing to the fact they can integrate and summarise different layers of data in an easy and comprehensible way. A detailed review and comparison of methods for discovering cancer drivers can be found in [140].

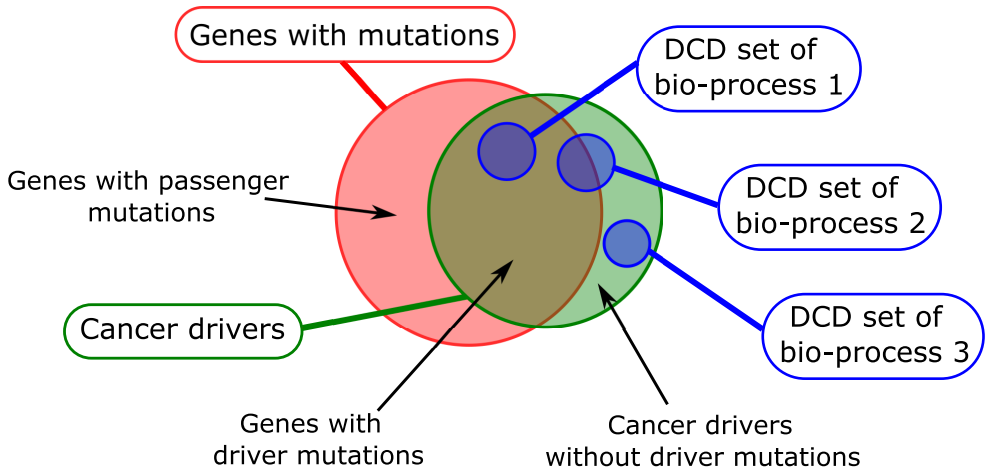


Figure 5.1: **Graphical representation of the overlap between different categories of cancer drivers (adapted from [140]).** Genes with driver mutations are cancer drivers. Recent findings have shown that certain genes without driver mutations can also drive cancer progression. Genes driving biological processes that lead to cancer (i.e. ***dynamic cancer drivers***, or DCDs) are not necessarily mutated. Additionally, different biological core processes (shortened to "bio-process" in the figure) are driven by different sets of genes.

The fundamental hypothesis of the proposed method is that "incorporating the dynamic aspect of biological processes driving cancer progression enables the detection of ***cancer drivers*** that are not identifiable by methods based on data that neglect dynamics" because neglecting dynamics during the inference process reduces the spectrum of cancer drivers that can be detected from observational data. Throughout this chapter, a gene driving one or more core processes in cancer progression is called a ***dynamic cancer driver*** (DCD). Genes driving such processes may not be mutated, as not all cancer drivers harbour driver mutations (see Fig. 5.1, and see [140] for a review). Moreover, anomalies in biological processes that lead to cancer development may occur at different stages of the cancer, and thus

the set of cancer drivers may not be fixed throughout the progression of the disease.

The hypothesis mentioned above is based on the fact that mutational and network-based approaches infer cancer drivers without using the dynamic aspect of the disease. Cancer, as well as a large number of other diseases, is a collection of heterogeneous processes, and it is considered to be driven by irregularities (including mutations) at different molecular levels, such as genetic, epigenetic, and transcriptomic levels [214]. These irregularities alter normal biological processes. As a result, phenotypes linked to the disease are the consequences of various biopathological processes occurring within a complex network [152], evolving through time in a dynamic way.

Dynamic information (defined as information generated by a dynamic system) on certain biological processes can be obtained directly from time series datasets (e.g. [179, 77]). However, reliable disease-related time series datasets are rare in practice, as they are costly and technically difficult to obtain. Moreover, disease progression is influenced by a wide range of factors that cannot be completely replicated artificially. Additionally, most time series datasets for biological processes are constructed from the results of controlled experiments in laboratories, and usually they can accurately represent only a few processes in the progression of the disease at a time. For these reasons, the vast majority of cancer-related datasets are obtained from cross-sectional studies, where data (known as cross-sectional or static data) are collected at a single instant during the course of the disease.

Capturing useful dynamical information from cross-sectional cancer data for discovering cancer drivers is the key step for exploiting dynamical information in static data. The *pseudotime* concept, originally developed for single-cell data, is an appropriate candidate for extracting dynamical information from biological processes. The basic idea of *pseudotime* is that, given a collection of heterogeneous samples (usually single cells) dispersed along a biological transition (a biological progression), each sample can be scored to reflect its stage in the progression [89]. Recent studies have demonstrated that, owing to its intrinsically temporal nature, *pseudotime* can successfully extract aspects of the underlying dynamics of a biological process from both bulk and single-cell datasets [84, 85, 86].

Three fundamental aspects should be considered during the process of inferring cancer drivers. First, cancer evolves over time [215, 216]; therefore, it is a collection of dynamic processes, and cancer drivers are actually driving one or more processes during cancer progression. Second, genomic aberrations that drive cancer progression are not limited to gene mutations [217, 75]; thus, inferring cancer drivers based on the gene mutational burden ignores information from other essential aspects of cancer evolution. Third, based on the definition of a cancer driver as a gene that promotes the appearance and progression of cancer, a cancer driver gene is a component in the causes of cancer development. Consequently, the causal nature of the driver-cancer relationship should be considered during the inference process. To overcome the limitations of current mutational and network-based methods for inferring cancer drivers while fulfilling the three fundamental aspects discussed above, a ***dynamic cancer driver*** inference approach was developed in this thesis.

The proposed method takes gene expression data from cross-sectional studies as well as a covariate that reasonably modulates (in the sense described in [84]) the *pseudotemporal* progression of one relevant process occurring during cancer development. If no *pseudotime* is provided, the method relies on *PhenoPath* [218, 84] to find a *pseudotime* score to order the samples following the trajectory encoded by the covariate. The *pseudotime* (either provided or inferred by using *PhenoPath*) and the covariate provided were used to find a critical turning point in the trajectory along the *pseudotime*. This critical point is named the "event" (further details on this are discussed in section 5.2.3). In either case, the proposed method requires a trajectory covariate.

The proposed hypothesise is that the causal relationship between a driver gene and cancer development induces a significant deviation (also referred to as *causal impact*) in a core process from normal to carcinogenic after the "event". This *causal impact* can be detected in the gene expression of such a driver gene. Thus, the *causal impact* of each "relevant" gene after the "event" is assessed (more details are given in section 5.2). If a gene expression reflects a significant *causal impact* after the "event", the proposed method considers the corresponding gene a cancer driver.

The proposed ***dynamic cancer driver*** approach was applied to a single-cell RNA sequencing dataset (NCBI GEO database, accession GSE75688) [219] and *The Cancer*

Genome Atlas Breast Cancer (TCGA BRCA) dataset [69] (the details are in sections 5.3.1 and 5.3.4, respectively). The proposed method's ability to expand current cancer driver catalogues as well as its inference power were assessed. The Cancer Gene Census (CGC) [220] was used as a conservative approximation of the ground truth because the CGC catalogue is still expanding, and thus non-CGC genes discovered by the proposed method should not be ruled out as drivers. Additionally, the TCGA-BRCA dataset was used to compare the proposed method to seven popular methods for inferring cancer drivers. The results showed that the proposed method discovered more CGC genes with better precision than the other methods, thus demonstrating its superior performance.

The experiments suggest that the proposed method can identify cancer driver genes for different trajectories from the same dataset. Unlike other methods that can only identify a global (static) set of cancer drivers, the proposed method discovers ***dynamic cancer drivers*** (i.e. genes driving one or more core processes in cancer progression). The results also suggest that the temporal information not only improves the performance of cancer driver inference, but also allows the genes driving specific processes throughout cancer development to be identified.

5.2 Methods

5.2.1 Problem definition and method overview

In the proposed method, a ***dynamic cancer driver*** is defined as a gene driving one (or more) significant biological processes throughout cancer progression. Considering this definition, the proposed method aims to identify cancer drivers by incorporating the dynamical information of cancer development into the inference process. The following two aspects are considered in the proposed inference method:

1. If a biological process is causing cancer to develop, at some point during its progression there was a critical event that caused significant disturbances, deviating the process from normal to carcinogenic. Although, a biological process is generally not disruptive, the proposed method includes the hypothesis that there is a point in the process at which the transition from normal to cancerous becomes irreversible.

2. If a gene is driving a cancer-related biological transition, it should cause significant changes that can be detected in its own *pseudotime*-ordered gene expression (i.e. the *pseudotime* series of the expression of the gene; see section 5.2.3.1). The abnormal gene expression is used as an indicator of the alterations in the behaviour of genes because the analysis of differentially expressed genes as indicators of cancer is a common and widely accepted approach.

Dynamical information required for the causal inference of cancer drivers depends on the process selected for study. Most cancer datasets contain cross-sectional data, and thus dynamical information is not readily available. To extract the underlying dynamics of a process from cross-sectional datasets, a covariate encoding significant aspects of the trajectory of interest and the *pseudotime* concept were used. Additionally, the *PhenoPath* framework [84, 218] was used for scoring each sample to reflect its relative position along the biological process (or trajectory). The *PhenoPath* framework was selected for *pseudotime* scoring because of its confirmed ability to extract dynamical information from both single-cell and bulk data [84].

The proposed method uses the *pseudotime score* to order the samples. The *pseudotime series* of a gene g is denominated as the ordered gene expression in which an observation g_t denotes the observation at sample t, and $t \in \{1, 2, ..., n\}$ indexes the samples in *pseudotime* order (this is further discussion in section 5.2.3). The proposed method can be summarised according to three phases:

1. Network-based selection of putative drivers (section 5.2.2): a specific *protein-protein interaction* (PPI) network is created from the dataset. The most influential genes in this network are labelled as putative cancer drivers.

2. *Pseudotime* series modelling and event detection (section 5.2.3): a provided trajectory/path covariate and PPI genes in the dataset are used for *pseudotime* scoring. The term *pseudotime* order is used to refer to the order of the samples that are sorted in ascending order in terms of the score. Scoring is performed by using the *PhenoPath pseudotime* model proposed in [218]. The calculated *pseudotime* score is used to create the *pseudotime series* for both genes and the path covariate. Each sample following *pseudotime* order is evaluated to detect the "event".

3. Cancer driver identification based on *causal impact* (section 5.2.4): the significance of the changes in the *pseudotime series* (after the "event") are assessed. Causal relationships between genes and the selected process in cancer (encoded by the path covariate) are assessed following a procedure based on the *causal impact* approach proposed by Brodersen et al. [221].

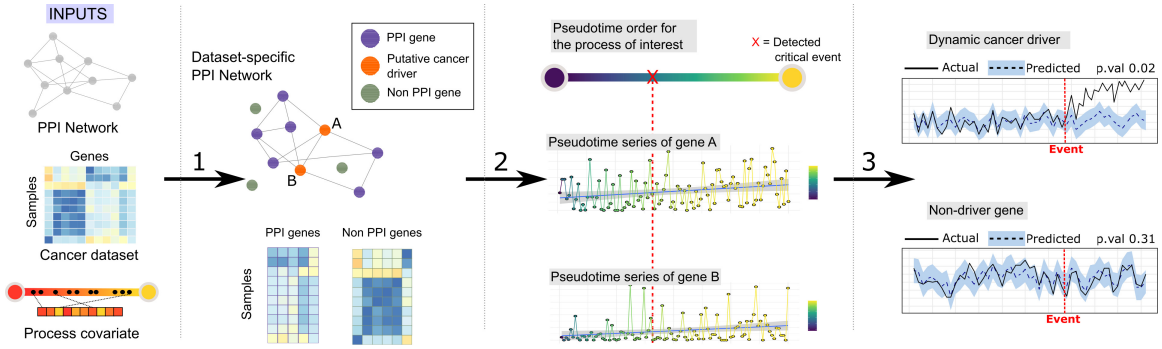


Figure 5.2: **Summary of the proposed method for identifying genes driving relevant processes in cancer development.** (1) Starting with a gene expression dataset, the proposed method infers a dataset-specific protein-protein interaction (PPI) network. The dataset is divided into PPI and non-PPI genes. Genes with the most interactions in the specific PPI network are considered putative drivers. (2) PPI genes and a trajectory covariate are used for inferring a pseudotime order that follows the process of interest encoded by the path covariate. Ordered samples are assessed to determine the most significant change in the process or the critical event. All gene expressions in the dataset are ordered to create pseudotime series. (3) The pseudotime series of the genes are used for creating a counterfactual model. *Causal impact* is calculated as the difference between the counterfactual and observed gene expression. If a significant causal effect is detected, the gene is labelled as a ***dynamic cancer driver***. If the predicted and actual gene expression values are not significantly different, the gene is labelled as a non-cancer driver.

5.2.2 Phase 1: Network-based selection of putative driver genes

Starting with a matrix $\mathbf{G} \in \mathcal{R}^{n \times m}$ containing the gene expressions from n samples and m genes, the proposed method performs a dimensional reduction $\mathbf{G}^{n \times m} \rightarrow \mathbf{G}^{n \times k}$ (with $k \leq m$) by removing any gene not expressed in at least 20% of the samples. The directed PPI network provided by [222] is used as input to classify the remaining genes as "PPI genes" or "non PPI genes". This network consists of 34814 gene-gene interactions from 6339 different genes. A data-specific PPI network is created from $\mathbf{G}$ as the PPI sub-network formed by the genes present in the dimensionally reduced dataset (see Fig. 5.2, item 1).

The proposed method considers a gene a *putative driver* if such a gene: (i) belongs to the data-specific PPI network inferred and (ii) the degree of its corresponding node

(where "degree" is defined as the total number of interactions of the gene in the inferred PPI network) is in the top 40% of the node degree distribution. Although this parameter can be adapted, the experiments showed that a suitable threshold stays between the top 50% and the top 25%.

5.2.3 Phase 2: Pseudotime series modelling and event detection

5.2.3.1 Pseudotime series

As cancer progression is assumed to be a collection of heterogeneous processes, there are several relevant trajectories that can be employed in the dynamic causal inference of cancer drivers. The definition of a process of interest is fulfilled by using a provided trajectory covariate P. It is assumed that the provided covariate P encodes significant aspects of the latent progression of the samples (e.g. patients, tumours, cells, etc). The assumption that such a covariate exists is reasonable because *pseudotime* order usually reflects the progression associated with prior factors of interest [84]. In general, a numeric vector (e.g. gene expression of a biomarker) or categorical value (e.g. estrogen-receptor, or ER, status in breast cancer) can be used as a suitable covariate.

Given an expression matrix of all "PPI genes" in the dataset-specific PPI network (obtained in phase 1, see Fig. 5.2) and a path covariate P, the proposed method utilises the *pseudotime* concept for scoring each sample to reflect its relative position along a process. In the proposed method, P is considered a numerical (continuous) covariate. To be valid, the selected covariate must have a unique value matching each sample in the dataset. If no *pseudotime* order is provided, the proposed method relies on *PhenoPath* [218] for scoring each sample. The dataset samples are then reordered in *pseudotime* order. Consequently, each column in the dataset is now considered a *pseudotime series* (Fig. 5.2, item 2).

5.2.3.2 Event detection

To detect the critical event, the *pseudotime series* of the path covariate is tested to determine the sample in which the most significant change is exhibited (i.e. the "event"). The proposed method orders the path covariate P by using the *pseudotime* score (as explained in section 5.2.3.1) to create its *pseudotime series*. It is assumed the "event" is not in the first five samples nor the last five samples (in *pseudotime* order). This is based on the further assumption that the provided dataset contains

samples that are reasonably dispersed along the trajectory. This requirement may not be fulfilled when using a dataset that includes only samples of a very limited region of the transition, which would not adequately reflect the progression of the biological process.

For the event detection, a modified version of *Causal Impact* is used [221] in the proposed method. To assess whether changes in a response variable are due to a treatment, *Causal Impact* requires a time series of the response variable and a contemporaneous time series that is unaffected by the treatment. *Causal Impact* uses the unaffected time series as a reference for counterfactual estimations.

In the proposed approach, the path covariate P is utilised as a proxy for a response variable, as its *pseudotime series* is expected to encode the changes due to the cancer of a biological process of interest. Changes in a ***dynamic cancer driver*** are expected to be similar to the changes encoded by the path covariate. For this reason, and because the putative drivers belong to the PPI network, "PPI genes" were ruled out as suitable references for the counterfactual estimations required by *Causal Impact*, and only "non PPI" are used genes as the references. The "event" identification is performed as follows:

a. Find the "non-PPI" gene having the largest *Pearson* correlation with the path covariate P.

b. Starting from the fifth sample (in *pseudotime* order), set the current sample as the "plausible event".

c. Use the data from all samples before the current "plausible event" (i.e. from sample 1 to the current "plausible event") to fit a regression model expressed as

$$\hat{P}_t = \mu_t + \delta_t + \beta \hat{x}_t + \varepsilon_t \quad \text{with} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{5.1}$$

where $\hat{P}_t$ is the predicted value of P at *pseudotime* t, $\hat{x}_t$ is the value at t of the *pseudotime series* of the "non-PPI gene" selected in step **a**, μ_t is the current local

linear trend level and δ_t is its respective slope, β is a regression coefficient, and ε_t is the error term.

d. Use the model obtained in **c** to obtain the predicted values of P for all samples from the current "plausible event" onwards.

e. Check whether the differences between the predicted and observed values are significant and assess if those differences are due to cancer. This is done by retrieving the *relative causal impact* for each "plausible event", which is calculated as

$$\text{relative causal impact} = \frac{\sum_{\text{post-period}} P_t}{\sum_{\text{post-period}} \hat{P}_t} - 1, \tag{5.2}$$

where "post-period" refers to the *pseudotime* values from the current "plausible event" onwards, P_t is the actual value of P at *pseudotime* t, and $\hat{P}_t$ is the predicted contemporaneous value (a detailed explanation of *Causal Impact* in our method is presented in 5.2.4).

f. Label the "plausible event" with the largest (significant) *relative causal impact* as the identified "event".

5.2.4 Phase 3: Cancer driver identification based on causal impact

Given a causal relationship between two variables A $\rightarrow$ B (read as "A causes B") that is ongoing from an arbitrary time t_e, the *causal impact* on B (the response variable) due to A (the cause) is described as the potential change (after time t_e) in B induced by A. In other words, the *causal impact* on B is the difference between the observed values of B after t_e and the value that B would have had if A had not existed [221].

In the proposed method, it is assumed that the causal effect that a cancer driver has on a biological process is reflected in the *pseudotime series* of such a driver. Consequently, the *causal impact* of a driver gene is defined as the difference between the values of the (observed) gene expression due to cancer and the (unobserved) gene expression of the same gene if cancer had not occurred. In the proposed method, t_e is defined as the index of the sample (in *pseudotime* order) for which

the most critical change occurred (i.e. the "event"). Furthermore, the response variables in the proposed approach are the *pseudotime series* of the putative drivers (section 5.2.2). The "event" t_e is identified according to the procedure explained in section 5.2.3.2.

To obtain the unobserved gene expression of a putative driver, a counterfactual model of the respective *pseudotime series* is required. Such a counterfactual model can be achieved by using another *pseudotime series* as a contemporaneous covariate to tune a regression model [223]. The "non-PPI gene" with the largest *Pearson* correlation was selected as the covariate for the putative driver. Covariates from the group of "non-PPI genes" were then selected, as they are outside the PPI network and it was hypothesised that changes in "PPI genes" after the "event" do not induce changes in the covariates.

The process used for obtaining counterfactual estimations of a putative cancer driver is as follows: (i) determine the sample representing the most significant change along the path of interest, or the "event" (as shown in section 5.2.3.2), (ii) find a counterfactual model of the putative driver using its data before the "event" (section 5.2.4.1), and (iii) use the model from (ii) to obtain the counterfactual data after the "event".

5.2.4.1 Counterfactual model

A state space representation of time series data is used as the counterfactual model to estimate the *pseudotime series* of the putative drivers if cancer had not happened. State space representation is a widely accepted approach for modelling time series and dynamic systems. It relates the observed values of the variables in a system to the latent states of those variables (see [224] for a comprehensive explanation).

In the proposed ***dynamic cancer driver*** method, given the expression profile of a putative driver gene $g \in \mathbf{G}$, its corresponding *pseudotime series* is modelled using the adapted version [225] of the state space representation model [221, 225, 226] expressed as

$$g_t = Z_t^T \alpha_t + \varepsilon_t \quad \text{with} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{5.3}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \quad \text{with} \quad \eta_t \sim \mathcal{N}(0, Q_t), \tag{5.4}$$

where μ_t is the current local linear trend level and δ_t is its respective slope, x_t (used as a covariate) is the contemporaneous value of the "non-PPI gene" having the largest *Pearson* correlation with g, β is a regression coefficient, and ε_t is a scalar observation error term following a normal distribution $\mathcal{N}(0, \sigma_t^2)$. The data before the "event" (in *pseudotime order*) is used to fit the state space model, which is used to obtain counterfactual estimations (from the "event" on). Equation 5.3 is known as the *observation* equation and it relates g_t (an observed value of the *pseudotime series* of gene g) to the vector of latent states $\alpha_t = [\mu_t \quad \delta_t \quad 1]^T$ through the *output vector* $Z_t^T = [1 \quad 1 \quad \beta x_t]$. Equation 5.4 is the *transition* equation and it describes the evolution of the latent states (α_t) over time; T_t is a squared transition matrix and $R_t \eta_t$ is the error term.

Counterfactual estimations $\hat{g}_t$ are achieved by evaluating equations 5.3 and 5.4 for all $t \geq t_e$ (from the "event" onwards). These estimations are used for calculating the *causal impact* reflected in the *pseudotime series* of the putative driver:

$$\text{Causal_Impact}_t := g_t - \hat{g}_t \qquad \forall \quad t_e \leq t \leq n. \tag{5.5}$$

In the proposed method, counterfactual modelling and the evaluation of the significance of the *causal impact* are performed by using the *CausalImpact* function in the *Causal Impact* R package [221].

5.2.4.2 Discovery of dynamic cancer drivers

In this step, putative drivers with a significant *causal impact* detected from its own *pseudotime* series are considered ***dynamic cancer drivers*** (Fig. 5.2, item 3). A *causal impact* of a putative driver is considered significant if the causal effect calculated by using the observed and predicted gene expression values after t_e (the "event") is statistically significant (*CausalImpact* p.val < 0.05).

5.3 Results

5.3.1 Dynamic cancer drivers identified from single-cell data

The proposed method was applied to the single-cell RNA sequencing data from the NCBI GEO database, accession GSE75688 [219]. This dataset contains gene ex-

pressions from 11 breast cancer patients (515 single-cell sequencing data points) with distinct molecular subtypes: estrogen receptor positive (ER+), human epidermal growth factor receptor 2 positive (HER2+), double positive (ER+ and HER2+), and triple-negative breast cancer (TNBC). The inference was performed by using 317 out of the 515 cells from the dataset, corresponding to all tumour cells provided. After removing low expressed genes, the dataset contained the gene expressions of 9551 genes.

The samples were ordered following two different *pseudotimes* calculated by using HER2 and VIM expressions as path covariates. HER2 was employed as the trajectory covariate for generic breast cancer progression because it is a well-recognised oncogene and its dysregulation influences breast cancer development. Similarly, VIM was used as the trajectory covariate for the *epithelial-to-mesenchymal* (*EMT*) process because of the fact that VIM is a well-recognised *EMT marker*. As a result, the proposed method identified two sets of drivers, one from each *pseudotime* order. For simplicity, these *pseudotime* orders are referred to as "HER2time(SC)" and "VIMtime(SC)" hereafter.

Table 5.1: **Summary of inferred dynamic cancer drivers from single-cell data when using HER2 and VIM as process covariates.** The sizes of the different sets of drivers were inferred from each pseudotime. "Combined(SC)" shows the number in the union set.

	HER2time(SC)	VIMtime(SC)	Combined(SC)
Total inferred	604	545	934
CGC genes	98	93	155
% of CGC genes	16.225	17.064	16.595

In this dataset, a total of 604 genes were inferred as drivers during generic breast cancer progression (as encoded by "HER2time(SC)"), while a total of 545 genes were inferred as drivers of *EMT* (as encoded by "VIMtime(SC)") (refer to Table 5.1). The proposed method was able to identify 98 and 93 experimentally confirmed cancer drivers listed by the *Cancer Gene Census* (v94, n=719) [220] (also called CGC genes) from "HER2time(SC)" and "VIMtime(SC)", respectively. For further analysis, these discoveries were combined into a single list, resulting in 934 inferred genes. A total of 155 CGC genes were identified (see Table 5.1).

As expected, in the experiments different set of genes were identified as drivers of each biological transition of interest. The 545 ***dynamic cancer drivers*** inferred for

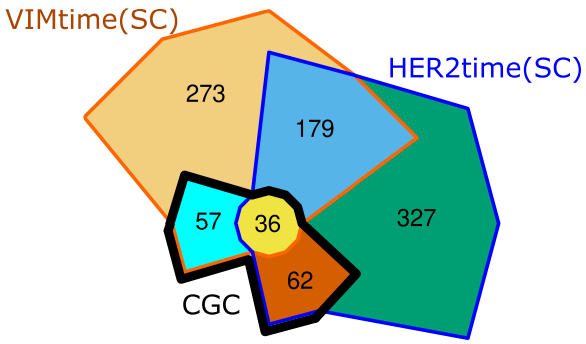


Figure 5.3: **Chow-Ruskey Venn diagram of discoveries of proposed method when exploring cancer drivers following the "VIMtime(SC)" and "HER2time(SC)" *pseudotime* orders**. The 155 CGC genes detected by the proposed method consist of 57 discovered from "VIMtime(SC)", 62 discovered from "HER2time(SC)", and 36 detected in both trajectories. The novel inferred drivers are distributed as follows: 273 genes were inferred by following "VIMtime(SC)", 327 genes were inferred by following "HER2time(SC)", and 179 genes were inferred to drive both trajectories. The figure was made using Intervene [207] (https://intervene.shinyapps.io/intervene/).

EMT as encoded by "VIMtime(SC)" and the 604 inferred as encoded by "HER2time(SC)" are distributed as follows: 36 CGC genes were detected as drivers in both trajectories, 57 CGC genes were detected by following (exclusively) "VIMtime(SC)", 62 CGC genes were detected driving cancer progression following (exclusively) "HER2time(SC)", 179 non-CGC genes were detected in both trajectories, 273 non-CGC genes were detected from "VIMtime(SC)", and 327 non-CGC genes were detected from "HER2time(SC)" (see Fig. 5.3). Appendices B.3 and B.4 show the top 100 ***dynamic cancer drivers*** from "HER2time(SC)" and "VIMtime(SC)", respectively, ranked by relative causal impact in descending order. The full list of drivers inferred from the single-cell data set employed can be found at `https://github.com/AndresMCB/DynamicCancerDriver`.

5.3.2 Contribution of proposed method to expanding current cancer driver catalogues

The metric described in Dietlein et al. [209] was adopted to assess the ability of the proposed method to contribute to the expansion of current cancer driver catalogues. The 155 CGC genes discovered as drivers by the proposed method (i.e. the union set of the "VIMtime(SC)" and "HER2time(SC)" inferences) were compared to CGC genes reported in previous (comprehensive) studies for discovering cancer driver genes. In this approach, the ability of the proposed method to ex-

pand previous catalogues is correlated with the number of additional CGC genes the method discovered. Additional CGC genes are defined as genes listed in the CGC that are inferred by the proposed method but not present in any of the cancer driver catalogues. A brief description of studies selected for this assessment is given below.

Six comprehensive studies providing catalogues of highly confident cancer drivers were selected. The first was Lawrence et al., which provided a catalogue of cancer driver genes obtained from 4742 tumour data points by using the MutSigCV suite [146]. The second was Martincorena et al. [227], which used the dNdScv algorithm on 7664 tumours to assemble a list of cancer drivers. The third was Bailey et al. [14], which created a cancer driver catalogue by applying a combination of 26 different computational tools to infer driver genes for 9423 tumours. The fourth was Priestley et al. [15], which provided a cancer driver mutation catalogue discovered from 2520 tumour-normal pairs of genomes from 2399 patients. The fifth was Dietlein et al. [209], which identified cancer drivers from 11873 tumour–normal pairs by using a novel method (MutPanning) based on the nucleotide context. Finally, the sixth was Rheinbay et al. [29], which provided a catalogue of coding and non-coding cancer drivers inferred by analysing 2658 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) study and The Cancer Genome Atlas (TCGA) [228].

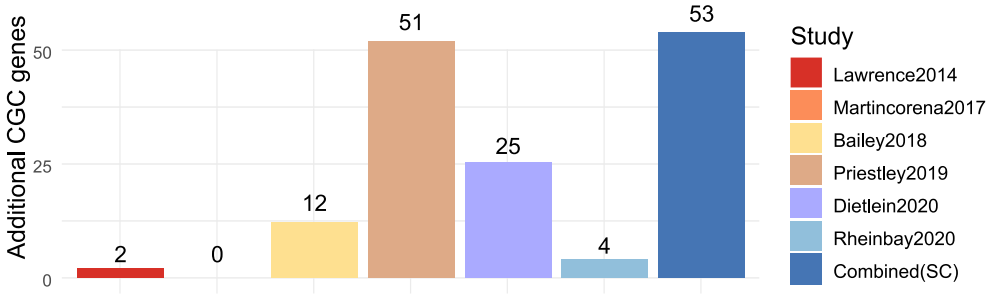


Figure 5.4: **Comparison between number of novel CGC genes inferred by the proposed method and those inferred by each of the cancer driver catalogues analysed.** The proposed method considers a CGC gene novel if such a gene does not belong to any of the analysed catalogues. Similarly, a CGC gene in a catalogue is considered novel if such a gene is not present in any of the other catalogues. The proposed method detected 53 novel CGC genes, suggesting that the method is useful for expanding current comprehensive cancer driver catalogues.

Out of the 155 CGC genes identified by the proposed method, 53 were not in any of the aforementioned catalogues. Similarly, the number of additional CGC genes

that can be found in each catalogue (i.e. CGC genes that are not in any of the other studies) were counted. The number of additional CGC genes from each catalogue were as follows: [15] 51, [209] 25, [14] 12, [29] 4, [146] 2, and [227] 0 (refer to Fig. 5.4). These results suggests that the proposed method's ability to detect additional cancer drivers is useful for expanding the findings of previous studies.

5.3.2.1 Stratification of inferred driver genes based on literature support

To further explore how the proposed method can contribute to the expansion of current cancer driver catalogues, a stratified list of ***dynamic cancer drivers*** was assembled. In the proposed catalogue, only the 215 genes discovered as cancer drivers in the intersection of both *pseudotime* orders (179 novel genes and 36 CGC genes; see Fig. 5.3) were considered, as they were relevant to both of the trajectories analysed.

Stratification indicates the confidence level of a driver gene. The 215 inferred genes were classified into the following four categories: **A** - the gene is a CGC gene, **B** - the gene has literature support as being influential in breast cancer, **C** - the gene has literature support as being influential in other cancers but not breast cancer, and **D** - there is no literature describing the gene as being influential in cancer. The inclusion criteria for the literature review were: (i) the paper explicitly refers to the gene as having a significant role in cancer, (ii) the publication is dated 2016 or later, and (iii) the study is published in a journal with an impact factor of 5.1 or higher (as indicated by the *web of knowledge*, website visited in August 2021).

Out of the 215 classified genes, 36 were CGC genes (A), 57 genes have been related to breast cancer (B), 69 genes had literature support regarding other cancers (C), and 53 genes had no literature support as of yet. For example, the receptor tyrosine kinase AXL, inferred by the proposed method as the gene with the largest *relative causal impact*, has been found to stimulate the phosphorylation of a network of focal adhesion (FA) proteins, which promote metastasis in breast cancer cells [229]. The same gene has been considered a potential target for cancer therapy because its knockdown is related to slower growth and increased radiosensitivity in breast cancer tumours [230]. The *relative causal impact* was used for ranking the discoveries because it is the percentage version of the *causal impact*, and thus it eliminates the bias towards highly expressed genes.

Genes in the top 10 of cancer drivers inferred by the proposed method (ranked by the *relative causal impact*) were distributed as follows: one gene has been found to be a confirmed cancer driver (IRF4 ranked in top 2), six genes (AXL, IGF1R, STC2, MBD3, SATB1, and ENO2) had support from experimental studies, and three genes had no literature support as of yet. IGF1R (top 5) has been found to co-regulate tumorigenesis and cancer progression in breast cancer [231]. STC2 (top 3) promotes *EMT* progression in colorectal cancer, inducing tumorigenesis [232]. MBD3 (top 4) plays a significant role in the *EMT* process and tumour metastasis [233]. SATB1 (top 7) plays a critical role in relevant pathways conferring anti-tumour immunity [234]. Lastly, the knockdown of ENO2 (top 9) has been associated with a reduction in the survival rate of leukaemia cells [235]. The full stratified list can be found in Appendix B.5. The full list with references from the literature can be found online in the *Supplementary* folder at https://github.com/AndresMCB/DynamicCancerDriver.

5.3.3 Enrichment analysis of dynamic cancer drivers inferred from single-cell data

An enrichment analysis (GO biological processes 2021) was performed to verify how significantly the cancer drivers discovered by the proposed method are related to processes in cancer. Additionally, the gene-disease associations among the inferred drivers were explored by inspecting the overlap between the inferred genes and the human gene-disease associations (GDAs) provided by DisGeNET [236].

The analyses showed that a significant number of the discovered drivers are strongly related to cancer. The top 5 GO biological processes and the top 5 GDA enriched terms for VIMtime(SC) (ranked by p-value) are shown in Table 5.2. The top 10 enriched terms for each of the enrichment analyses are provided in Appendix B.6 ("GO terms HER2time(SC)"), Appendix B.7 ("GO terms VIMtime(SC)"), Appendix B.8 ("GDAs HER2Time(SC)"), and Appendix B.9 ("GDAs VIMtime(SC)"). An extended visual representation (top 10) for both GO terms and GDA terms can be found in Appendix B (Figures B.1 and B.2). The full results of the analyses (including the list of enriched genes) can be found at https://github.com/AndresMCB/DynamicCancerDriver. The cancer drivers discovered by following the *EMT* process are in general related to generic processes in cancer (e.g. carcinogenesis in top

1 and neoplasm metastasis in top 5 of the DGAs), while the cancer drivers inferred by following "HER2time(SC)" are more closely related to breast cancer.

Table 5.2: **Top 5 GO biological processes (left) and DisGeNET DGAs (right) enriched terms among cancer drivers inferred from VIMtime(SC)**

GO biological process term	**P-value**	**Disease-gene association term (DisGeNET)**	**P-value**
Positive regulation of transcription, DNA-templated (GO:0045893)	6.05E-41	Carcinogenesis	1.39E-55
Regulation of transcription by RNA polymerase II (GO:0006357)	2.28E-29	Malignant neoplasm of breast	1.64E-54
Positive regulation of transcription by RNA polymerase II (GO:0045944)	4.79E-29	Breast carcinoma	2.31E-53
Negative regulation of transcription, DNA-templated (GO:0045892)	6.06E-29	Malignant neoplasm of prostate	3.05E-52
Regulation of transcription, DNA-templated (GO:0006355)	1.60E-27	Neoplasm metastasis	3.88E-52

The overlap between the inferred drivers and the top 3 GO biological terms (2021) resulting from the enrichment analysis are as follows. For drivers inferred by using "VIMtime(SC)", positive regulation of transcription, DNA-templated (GO:0045893) (125/545, p.val 6.05E-41), regulation of transcription by RNA polymerase II (GO:0006357) (154/545, p.val 2.28E-29), and positive regulation of transcription by RNA polymerase II (GO:0045944) (93/545, 4.79E-29). For drivers discovered by following "HER2time(SC)", positive regulation of transcription, DNA-templated (GO:0045893) (144/604, p.val 2.06E-49), positive regulation of transcription by RNA polymerase II (GO:0045944) (114/604, p.val 6.47E-40), and regulation of transcription by RNA polymerase II (GO:0006357) (173/604, p.val 1.01E-33).

Similarly, the overlap between the inferred set of drivers and DisGeNET cancer-associated genes includes a significant number of genes. The overlap between the inferred drivers and the top 3 GDA terms resulting from the DisGeNET analysis are as follows. For drivers inferred by using "VIMtime(SC)", carcinogenesis (273/545, p.val 1.39E-55), malignant neoplasm of breast (306/545, p.val 1.64E-54), and breast carcinoma (301/545, p.val 2.31E-53). For drivers discovered by following "HER2time(SC)", mammary neoplasm (244/604, p.val 1.71E-73), malignant

neoplasm of breast (353/604, p.val 4.62E-69), and breast carcinoma (348/604, p.val 4.75E-68). These results suggest that the inferred cancer driver genes are significantly related to cancer.

5.3.4 Discovery of cancer drivers from bulk data using dynamical information

Cross-sectional data, also known as bulk data, account for the majority of data from cancer patients. For this reason, significant coordinated efforts to create repositories hosting bulk data useful for research purposes have been made. For example, *The Cancer Genome Atlas (TCGA)* program (https://www.cancer.gov/tcga) provides a comprehensive collection of genomic, epigenomic, transcriptomic, and proteomic cancer data [69]. The availability of these datasets is one of the primary reasons most of the current methods for discovering cancer drivers are designed for bulk data.

To further assess the performance of the proposed method and to validate whether including dynamic causal inference procedures enhances its ability to identify cancer driver genes from bulk data, the proposed method was applied to the TCGA-BRCA cancer dataset. The data used for this analysis were downloaded using the *TCGAbiolinks* package [237]. A total of 752 primary solid tumour samples from 747 patients (the same patients analysed by Pham et al. [75]) were selected. The downloaded data consist of gene expression (raw counts) harmonised against GRCh38 (hg38). After removing low expressed genes, the dataset had 19351 genes. To enable a fair comparison with mutational based approaches, the cancer drivers discovered by the proposed method were ranked by the frequency of the single nucleotide variants (SNVs) in this dataset (the more the gene appears as mutated, the higher the rank). Mutational information was obtained from the Mutation Annotation Format (MAF) (MAF.hg38) data accessed through *TCGAbiolinks*.

This dataset was normalised to compensate for the effects within the samples and the cross sections between the samples. This normalisation was performed by using the *TCGAanalyze_Normalization* method in the *TCGAbiolinks* package, which uses certain procedures to adjust for guanine-cytosine content (GC-content) effects on the read counts [238]. The VIM and HER2 gene expressions after normalisation were used as path covariates to find the corresponding "VIMtime(Bulk)" and

"HER2time(Bulk)" *pseudotime* orders from this dataset. The COSMIC CGC (v94, n=719) was used as the reference set of CGC genes.

The performance of a method was assessed according to the total number of CGC genes it was able to identify. The more CGC genes the method recalls, the better the method. To facilitate the comparison, the performance of the benchmarked methods in the top 50, 100, 150, and 200 cancer drivers predicted by each method were examined. OncodriveCLUST, OncodriveFM, and NetSig utilise the corrected p-value to rank their discoveries. The genes inferred by ActiveDriver and DriverNet are ranked by p-value. DawnRank provides its own ranking scores. CBNA also provides its own ranking based on the mutation frequency of the genes. The results of the performance evaluation are summarised in Fig. 5.5.

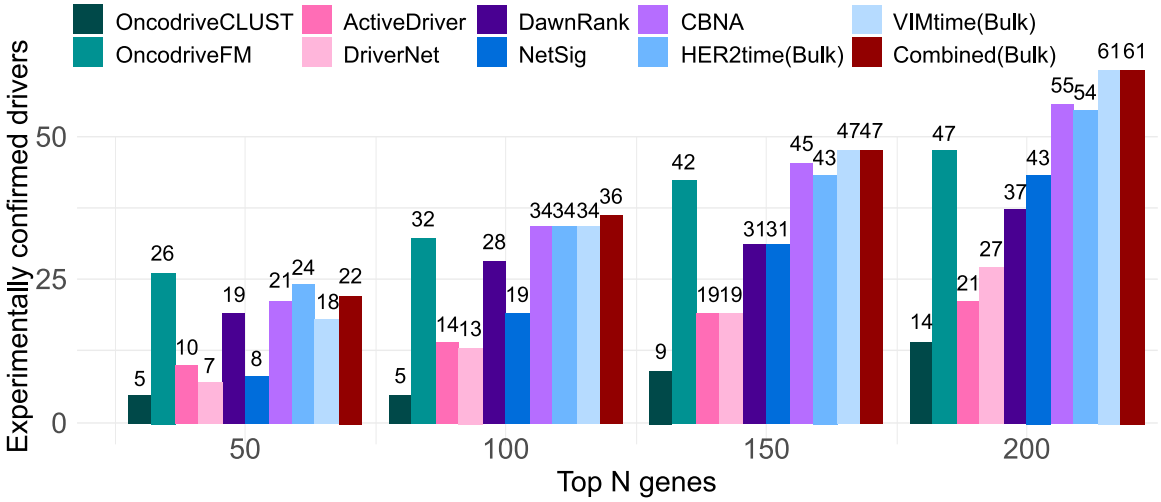


Figure 5.5: **Comparison of different cancer driver inference methods with bulk data.** The performance of the proposed method was benchmarked against seven other methods for discovering cancer drivers. CGC genes were used as a conservative approximation of the ground truth, as the full set of cancer driver genes still remains unknown. The results for the top N=50,100,150,200 genes following each method's rankings are presented. For a fair comparison with other mutational approaches, the inferred genes were ranked according to the mutation frequency in the TCGA-BRCA dataset.

The last three bars of each of the analysed top genes shown in Fig. 5.5 correspond to the performance of the proposed method for cancer drivers inferred from "HER2time(Bulk)", "VIMtime(Bulk)", and the two combined (union set). In general, the proposed method outperformed other methods when inferences were made following the *EMT* process ("VIMtime") and the two combined (union of the two inferred sets). The inference performance of the proposed method following "HER2time" was comparable to that of CBNA, and it outperformed the

other methods. These results suggest that the proposed method is able to identify mutated cancer drivers even when mutational information is not used during the inference process. These results also suggest that including information on the underlying dynamics (as well as causality) can enhance the proposed method's ability to discover cancer drivers, even in bulk data. The top 100 ***dynamic cancer drivers*** inferred from "HER2time(Bulk)" and the top 100 from "VIMtime(Bulk)" are shown in Appendix B (B.10 and B.11). The full list of ***dynamic cancer drivers*** discovered in the bulk data experiments can be found in the *Supplementary* folder at https://github.com/AndresMCB/DynamicCancerDriver.

5.3.5 Gene regulatory networks for dynamic cancer driver genes

The potential gene regulatory relationships between the ***dynamic cancer drivers*** (DCDs) discovered by the proposed method and the CGC genes that are marked as breast cancer drivers by the COSMIC project were explored. The regulatory relationships were retrieved from the top 200 inferred textbf*dynamic cancer drivers* by using HER2 as the path covariate (for both single-cell and bulk data) as well as the genes detected by the proposed method that were marked as breast cancer drivers in the COSMIC CGC database [220]. The regulatory relationships were extracted from http://www.grndb.com [239]. Only the regulatory relationships in the TCGA-BRCA dataset with high confidence were used. The resulting catalogue contained 93699 regulatory relationships. From this result, all regulatory relationships between the top 200 DCDs and the set of CGC breast cancer drivers (not restricted to the top 200 DCDs) detected by the proposed method (when using "HER2time(SC)" and "HER2time(Bulk)") were extracted.

Following this procedure, a total of 81 regulatory relationships were detected among the DCD genes discovered from the single-cell data. Similarly, a total of 56 regulatory relationships between DCD genes were detected from the bulk data. In both cases, widely accepted breast cancer drivers, such as BRCA1 and ESR1, were central nodes in the network (Fig. 5.6). The network analysis revealed several influential DCDs that are not curated CGC genes yet, such as E2F2 and THRA from the single-cell data (Fig. 5.6A) and SIN3A from the bulk data (Fig. 5.6B). E2F2 has been linked to tumorigenesis, metastasis, and poor overall survival in breast cancer [240]. THRA has been found to be correlated with high breast cancer mortality [241], and SIN3A mutations in breast cancer enhances cell proliferation [242]. The list of regulatory relationships used for plotting Fig. 5.6 can

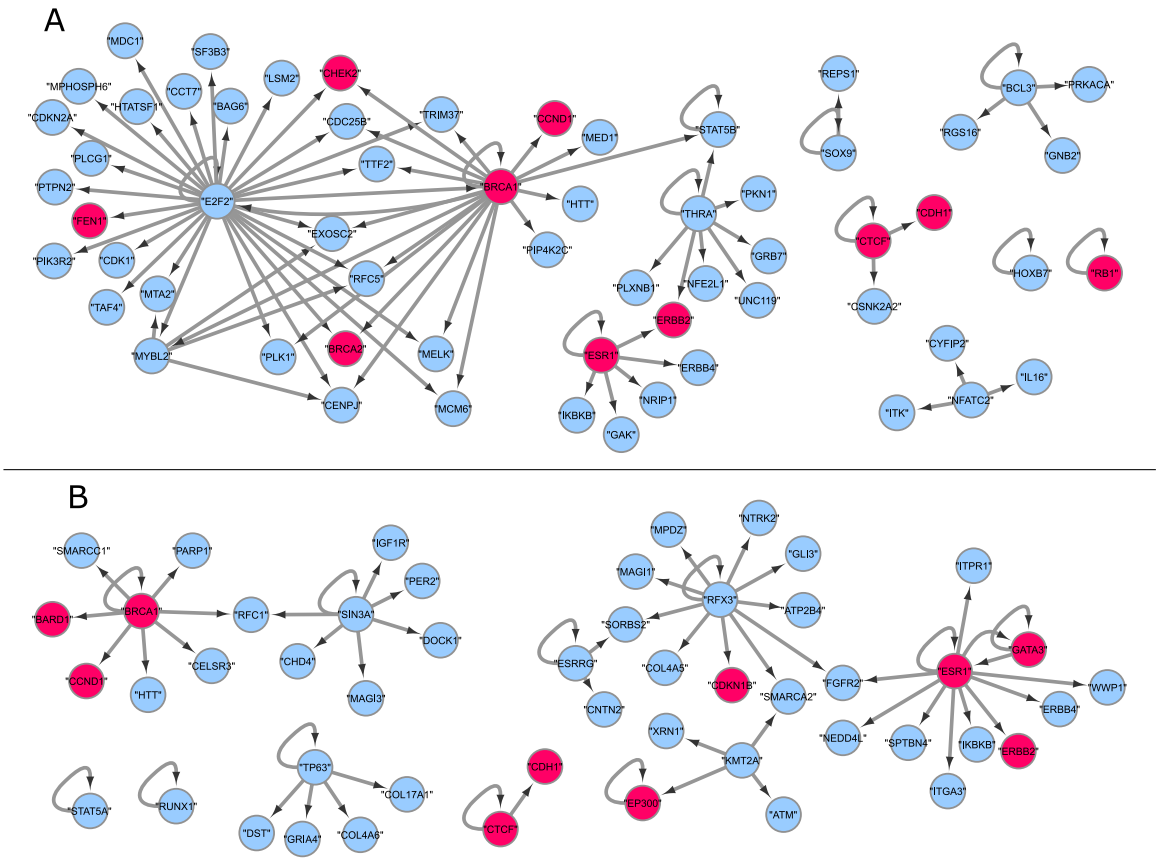


Figure 5.6: **Regulatory relationships among DCDs.** Gene regulatory networks showing genes in the union set of top 200 DCDs (blue) and all CGC breast cancer drivers identified as DCDs (magenta). A network analysis revealed novel cancer drivers that are highly influential in the network, such as E2F2 and THRA from the single-cell data (A) and SIN3A from the bulk data (B).

be found in Appendix B.12 ("DCD(HER2_SC) Reg. relationships") and Appendix B.13 ("DCD(HER2_Bulk) Reg. relationships").

5.3.6 Cancer drivers suitable for biological process covariates

Owing to the capability of the proposed method to use a gene as a covariate for orienting an inferred trajectory that fits a biological process of interest, the proposed method allows the use of well-recognised cancer drivers as path covariates to potentially discover DCDs for different cancer types/subtypes.

To further explore the proposed method's usability and stability when well-recognised cancer drivers are used as the path covariate, additional experiments were performed by using the ESR1 gene expression as the path covariate. ESR1 was selected

Table 5.3: **Summary of inferred sets of dynamic cancer drivers from single-cell data when using ESR1 as the process covariate.** The proposed method displays a similar behaviour when using a well-recognised breast cancer driver as the covariate to orient the pseudotime scoring process in comparison with the original experiments.

	ESR1time(SC)	HER2time(SC)	VIMtime(SC)
DCD inferred	500	604	545
CGC genes identified	83	98	93
% of CGC genes	16.6 %	16.255 %	17.064 %

because this gene is a curated breast cancer driver (CGC rank: tier 1). The results from the single-cell data revealed that the proposed method exhibited a consistent behaviour, as it identified a similar number of novel dynamic cancer drivers and CGC genes when using "ESR1time" in comparison with the original experiments ("HER2time" and "VIMtime"). Similarly, additional experiments using the bulk data revealed that the proposed method outperformed most of the benchmarked methods in detecting CGC genes when using "ESR1time". The results of the experiments that used ESR1 as the path covariate are summarised in Table 5.3 and Figure 5.7. The full list of DCDs inferred by using ESR1 as the process covariate can be found online at `github.com/AndresMCB/DynamicCancerDriver` in *"supplementary table 12 - dynamic cancer drivers ESR1time(SC)"* and *"supplementary table 13 - dynamic cancer drivers ESR1time(Bulk)"*.

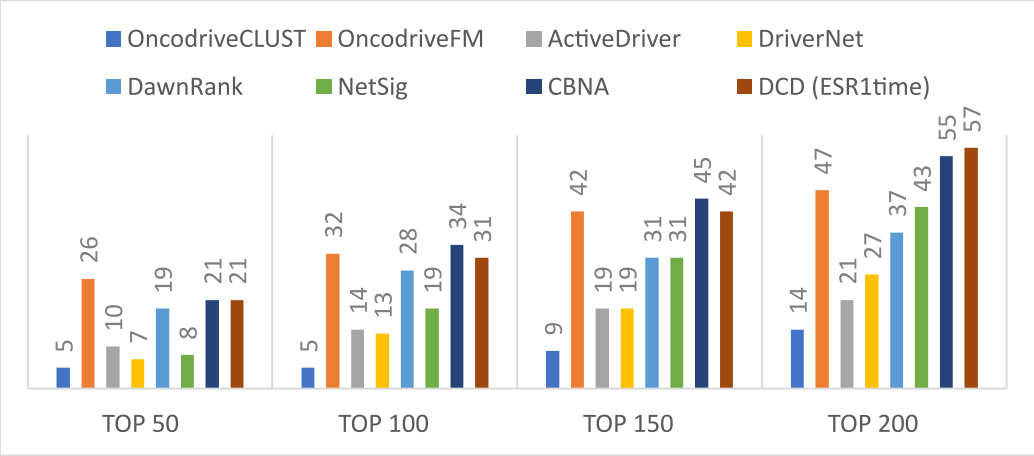


Figure 5.7: **Performance of the DCD method with ESR1 as a path covariate against various cancer driver inference methods for bulk data** The benchmarking process described in section 4.3.5 was replicated. The proposed method outperformed most methods and was the best method when including the top 200 DCDs. This result suggests that curated cancer drivers are suitable path covariates for the proposed method.

5.3.7 Differentially expressed gene (DEG) analysis of DCD genes inferred from bulk data

As an additional experiment, a differentially expressed gene (DEGs) analysis was performed to determine whether the detected DCDs were differentially expressed between normal and cancerous conditions. The DEG analysis was performed on the TCGA-BRCA dataset (as accessed in April 2021) by keeping all samples from normal tissues and primary tumours. This dataset had 1221 samples (113 normal samples and 1101 primary tumour samples). An R version of the dataset used can be found at http://4llab.net/Bioinformatics (in the "Datasets" section).

After removing low expressed genes (those not expressed in at least 20% of the samples), a total of 19505 genes were analysed. The TCGAanalyze_DEA function (fdr.cut = 0.01, logFC.cut = 1, method = "glmLRT") from the TCGAbiolinks package was used for the analysis. A total of 6156 genes were identified as differentially expressed. The full list of DEGs obtained from this analysis can be found online at https://github.com/AndresMCB/DynamicCancerDriver in *"supplementary table 14 - TCGA-BRCA DEGs"*.

The DEG analysis results were contrasted with the top 100 DCDs inferred from "HER2time(Bulk)" and the top 100 DCDs inferred from "VIMtime(Bulk)". The experiments showed that in both cases, a significant number of DCDs were not detected as DEGs. Specifically, 65 DCDs discovered by using "HER2time(Bulk)" (Appendix B.14) and 60 DCDs discovered by using "VIMtime(Bulk)" (Appendix B.15) were not retrieved as differentially expressed between the normal and cancerous conditions in the dataset. These results suggest that the gene expression changes of certain cancer drivers are subtle enough to prevent them from being detected by DEG analysis, and yet they can significantly alter the dynamics of a number of biological processes, causing cancer development.

Finally, the CGC genes in each top 100 that were not DEGs were filtered such that only genes that have been found to drive breast cancer were kept. When "HER2time(Bulk)" was used, the results showed that 8 out of the 26 CGC genes were curated breast cancer drivers, but they were not detected as DEGs. Similarly, an analysis of the top 100 DCDs when using "VIMtime(Bulk)" revealed that 7 out of the 24 CGC genes detected by the proposed method were curated breast cancer

Table 5.4: **Comparison between DEGs and DCDs from the TCGA-BRCA dataset.** A DEG analysis of normal tissue and primary tumour conditions was performed to determine if the discoveries of the proposed method are DEGs. The top 100 DCDs from HER2time and VIMtime were compared. The results suggest that a significant number of changes in gene expression between the normal and cancerous conditions are subtle enough to prevent certain cancer drivers from being detected as DEGs.

	not DEGs in top 100 DCD	CGC genes not DE in top 100 DCD	BRCA CD not DE in top 100 DCD
By using HER2time(Bulk)	65	26	8
By using VIM-time(Bulk)	60	24	7

drivers, but they were not detected as DEGs. These results support the hypothesis that the underlying dynamics of the disease provides insightful information that can improve the detection of cancer drivers. A summary of the DEG comparison is presented in Table 5.4.

5.3.8 DCD approach is useful for detecting drivers from different pseudotime methods

The proposed DCD approach uses *PhenoPath* for pseudotime order scoring, as it allows the inferred trajectory to be orientated to fit a biological process of interest using a covariate related to such a trajectory. Because of this capability, cancer drivers of different biological processes leading to cancer progression can be inferred. Other pseudotime inference methods lack this capability, which limits the ability of the proposed method to detect cancer drivers of different altered biological processes. However, the proposed method has proven to be useful for detecting cancer drivers from pseudotimes provided by external methods. Under these conditions, the path covariate is used only for detecting the critical event and not for orientating the trajectory to fit a biological process. As a result, the inferred DCD for different covariates (e.g. HER2, VIM) are expected to be similar.

To assess the above, further experiments were performed by using Monocle3 [243, 88] (a popular method for pseudotime inference) on the single-cell dataset GSE75688. For comparison purposes, HER2 and VIM gene expressions were used as covariates to detect the critical event seen from the trajectory inferred for Monocle3. The

	Monocle3 all DCDs	Monocle3 CGC genes	PhenoPath all DCDs	PhenoPath CGC genes
DCD (HER2)	687	120	604	98
DCD (VIM)	665	126	545	93
Jaccard similarity index	0.7468	0.7447	0.2301	0.2322

Jaccard similarity index (JI) was used to assess the similarity of the set of cancer drivers inferred when using VIM and the set of cancer drivers inferred when using HER2. The sets of cancer driver genes inferred when using HER2 and those inferred when using VIM were significantly similar for the pseudotime obtained by Monocle3 (JI = 0.7540). This result was not surprising, as the method inferred one "overall" trajectory from the dataset. In contrast, the JI for the sets of DCDs inferred by using PhenoPath pseudotime with covariates HER2 and VIM was significantly lower (JI = 0.2301), supporting the hypothesis that different biological processes have different sets of driver genes. These results are summarised in Table 4. The full list of DCDs inferred by using the Monocle3 pseudotime can be found in Appendix B.16 (event detection using HER2) and Appendix B.17 (event detection using VIM).

5.4 Discussion and conclusions

In this chapter, a new dynamic causal inference method for discovering cancer drivers was presented. The proposed method considers cancer drivers as genes driving one or more biological processes leading to cancer. Based on this consideration, the proposed method hypothesises that cancer driver genes are driving the dynamics of pathologically altered biological processes and that its causal influence in cancer development is reflected in its own gene expression.

To verify this hypothesis, the proposed method was used to perform a data transformation to unveil the underlying dynamics linked to one (or more) trajectories of interest and verify the possible causal link between putative cancer drivers and the progression of the biological trajectory. Furthermore, the sample that represented the most significant change along the biological transition was identified. That sample was used to determine the *causal impact* of the cancer driver. If such an impact was significant, the proposed method considered the gene a cancer driver.

In the proposed method, a PPI network is used to guide a knowledge-based feature selection that defines the putative drivers. However, hub genes obtained from hub identification methods and other methods for finding important genes in gene regulatory networks (see [244] for a comprehensive review of gene regulatory network methods for single-cell data) are also suitable putative cancer drivers in the proposed ***dynamic cancer drivers*** (DCD) framework. The causal inference method carefully explores such influential genes to identify DCDs by assessing the causal effects of the putative drivers on cancer, as reflected in the pseudotime-ordered gene expression of those putative drivers.

In the experiments, HER2 and VIM were used to demonstrate the proposed causal framework. However, the proposed DCD method allows the use of other known breast cancer drivers as well (e.g. ESR1). This capability can be extrapolated to known drivers of other types of cancer. As a result, the proposed DCD approach is also suitable for detecting drivers in other types/subtypes of cancer when a number of driver genes are known.

One interesting point about the proposed method is that it can detect cancer drivers that are not differentially expressed. The experimental results showed that 65 inferred DCDs (in the top 100 when using "HER2time(Bulk)") were not DEGs between the normal and cancerous conditions, including 8 CGC breast cancer drivers (i.e. PIK3CA, CDH1, NCOR1, MAP2K4, AKT1, CASP8, EP300, and CTCF). These results suggest that the gene expression changes of certain cancer drivers are subtle enough to prevent them from being detected by a DEG analysis, and yet they can significantly alter the dynamics of a number of biological processes, causing cancer development.

Additional experiments comparing *PhenoPath* (native) and an externally provided pseudotime (i.e. *Monocle3*, a popular method for *pseudotime* inference [243, 88]) showed that the proposed method can detect DCD genes from externally provided *pseudotimes*. *PhenoPath* allows the inferred trajectory to be orientated to fit a biological process of interest by using a covariate related to such a trajectory. Owing to this capability of *PhenoPath*, cancer drivers of different biological processes leading to cancer progression can be inferred. Other pseudotime inference methods lack this capability. As a result, in the experiments the DCD method could identify more novel cancer drivers from different path covariates when using *PhenoPath*,

supporting the hypothesis that different biological processes have different sets of driver genes.

Experiments with the data indicated the superior performance of the proposed method. The proposed method's ability to detect novel cancer drivers that can be used for expanding current cancer driver catalogues was assessed, and experimentally confirmed cancer drivers were identified. The analyses suggest that the ***dynamic cancer driver*** concept is useful for expanding the spectrum of current cancer driver catalogues. Additionally, the proposed method performed well for both single-cell and bulk data, outperforming all other methods analysed in terms of the number confirmed cancer drivers that were predicted for each method when using bulk data. The results of the tests suggest that using temporal information on biological processes significantly improves the efficiency of discovering cancer drivers. Additionally, the experiments suggest that certain cancer drivers significantly influence more than one biological trajectory in cancer progression.

Chapter 6

Identifying cooperative genes that cause cancer progression using dynamic causal inference

In the previous chapters, two dynamic causal inference methods for solving two of the sub-problems related to the dynamics of genetic relationships throughout the progression of cancer were presented. Specifically, these methods elucidated: (i) the structure of the gene regulatory networks that characterise biological processes in cancer (the ***PTC*** method), and (ii) the genes that drive the pathologically altered processes that cause cancer progression (the ***DCD*** method).

In this chapter, a novel causal method for identifying collaborative networks of cancer drivers based on dynamic system analysis is introduced. The method integrates the temporal dimension of the data throughout cancer progression and provides a way of testing for the causality of candidate genes in cancer.

The content of this chapter is based on the paper entitled "Identifying cooperative genes causing cancer progression with dynamic causal inference" [245], submitted to *Briefings in Functional Genomics*. A preprint version is available at https://biorxiv.org/cgi/content/short/2023.11.22.568367v1

6.1 Introduction

Cancer is one of the leading causes of death worldwide, accounting for one out of six deaths and nearly 18 million new cases in 2020 [3]. One key aspect of this disease is that treating it in its early stages sharply increases the likelihood of survival [2]. For this reason, understanding the dynamics of cancer (i.e. how it progresses) is crucial for successfully detecting and treating cancer. However, our current knowledge of the dynamics of cancer progression is far from complete, in part due to the challenges involved with the modelling of cancer at different stages [3].

Cancer genesis and development are linked to aberrations in genes commonly known as cancer drivers. It has been demonstrated that such aberrations provide a strategic evolutionary advantage to certain cells by altering critical biological processes in these cells. However, both genetic aberrations and their impacts on biological processes have not yet been fully characterised. Therefore, detecting cancer drivers and understanding their temporal behaviour are crucial tasks in cancer research. Most of the current approaches for inferring cancer drivers focus on gene mutations within cancer cells that consistently appear at a higher rate

than expected by chance [14, 246, 29]. In this paradigm, the discovery of cancer drivers relies on the frequency of the occurrence of such mutations (commonly called driver mutations).

Despite the success of current mutational approaches for discovering cancer drivers, it is now known that aberrations that cause cancer are not limited to mutations and they can appear at different levels, such as the epigenetic level [16, 17, 140]. Moreover, rare alterations (i.e. low-frequency events) contributing to cancer development are difficult to detect in data. As a result, significant effort has been committed to integrating additional information in order to compensate for the incompleteness in the approaches based on mutations. These new methods can be categorised as network-based methods (e.g. [73, 74, 75]) and methods integrating multi-omics (e.g. [247, 248]). Interesting reviews on these methods can be found in [244, 140, 153].

Most current methods for discovering cancer drivers do not consider the temporal information of the disease, nor do they include suitable procedures for testing the causal nature of cancer progression. Consequently, current methods provide few insights into the accurate modelling of cancer as a dynamic system. Furthermore, the dynamics of the causal relationships that occur during cancer progression are neglected. A lack of formal dynamic causal inference procedures results in a reduced capability of discovering cancer drivers from data.

The information required for the dynamic modelling of cancer can be obtained from time series data (e.g. [179, 77]). Time series are formed by collecting data from samples at different points in time during the progression of the disease. In practice, the collection of time series that describe biological processes involved in cancer is difficult and costly. Therefore, most of the available data correspond to snapshots taken from samples of cancer tissues (bulk data). Consequently, the vast majority of cancer driver inference methods have been developed for bulk data, which captures a minimal amount of information on the dynamics of cancer drivers.

One of the most promising approaches for extracting the underlying dynamics of biological processes relies on the *pseudotime* concept [84]. *Pseudotime* inference

methods commonly use the gene expression of samples at different stages of a biological transition to calculate a *pseudotime* score. This score aims to reflect the position of the sample in the progression of the biological transition. A large number of studies have shown that *pseudotime* scoring is intrinsically temporal. Furthermore, *pseudotime* has been shown to be useful for extracting part of the underlying dynamics of biological processes, even from bulk data [84, 85, 86, 87, 78].

By definition, a biological process involves the joint action of a specific set of genes, the expressions of which are highly regulated to follow a particular temporal sequence. Individual identification of cancer drivers (which is popular in most current methods) sidesteps this critical fact. To overcome this limitation, in this chapter an integrative method that models the causal relationships of cancer drivers influencing cancer progression (groups of driver genes rather than single drivers) is proposed. The method includes a temporal dimension to the data and formal procedures for testing causality in time-dependant data.

In the proposed method, to explore the dynamic aspect of cancer drivers, a temporal dimension to the data is included using the concept of *pseudotime*. Additionally, cancer drivers are identified by assessing the causal influence that driver genes have on the gene expression alterations occurring during cancer development. The proposed method explores the dynamics of cancer and the causal nature of driver genes by modelling the disease as a set of *causal kinetic models* (*CKMs*) [249]. A *CKM* is an extension of the *structural causal models* to the setting of *ordinary differential equations* (*ODEs*). Specifically, *CKMs* are used to model causal relationships involving cancer drivers as a collection of causal *ODEs*. The implementation of the method is discussed in section 6.2.

The method is applied to the TCGA-BRCA cancer dataset [69] and the single-cell RNA sequencing dataset in the NCBI GEO database (accession GSE75688) [219]. The results suggest that the proposed method is capable of systematically identifying cancer driver genes (hyper-geometric test p-value < 0.05) based on the dynamics of the interactions detected in the data. Additionally, the proposed method returns the set of dynamic models that are the most consistent with the observed data, providing insights into the underlying behaviour of biological processes in cancer as a dynamic system (in the sense that the system's current state depends

on the past). Furthermore, the proposed method can detect interactions and collaborations among the genes that cause cancer rather than focusing exclusively on the influence of single cancer driver genes on cancer development. As a result, the proposed method can elucidate the behaviour of genes that act together to cause cancer progression.

6.2 Methods

6.2.1 Problem definition

The premise of the proposed method is that genes driving cancer progression disrupt the balance of core biological processes, such as cellular division and apoptosis [250, 4, 6]. Because alterations in biological processes due to cancer drivers are progressive over time, the causal relationships involving driver genes can be considered a dynamic system. It is possible to consider each biological process as a dynamic system of genetic interactions. However, cancer itself is a highly complex dynamical system because it progresses over time (i.e. its future states depend on the past).

Such a dynamic system involves both the interaction and the collaboration of genes in the regulation of biological processes. Despite significant advances in genomics, there are still gaps in our understanding of how genes interact with each other. In particular, models of cancer as a dynamic system involving interactions and collaborations between genes have not yet been fully explored. In this chapter, a causal model of cancer (as a dynamical system) that takes into account the collaborations and interactions of genes causing cancer progression is presented.

6.2.1.1 Assumptions

In the proposed method, genes are considered random variables and the data correspond to the gene expression of m genes and n samples represented by a matrix $\mathbf{G} \in \mathbb{R}^{n \times m}$. The existence of a subset of genes in $\mathbf{G}$ that cause disruptions in one or more core biological processes during cancer progression is assumed. To assess causality, a target variable $y \in \mathbb{R}^{n \times 1}$ is also required. It is expected that the changes in the target variable are due to the influence of cancer drivers. Consequently, the target variable encodes significant information on cancer progression. Suitable tar-

get variables for the proposed method include relevant genes for cancer, such as well-recognised biomarker genes.

It is hypothesised that the progression of cancer (encoded by the target variable y) can be modelled as a *causal kinetic model*. A *causal kinetic model* (originally proposed by [249]) extends the concept of *structural causal models* (*SCMs*), a popular approach to causal modelling, to the setting of *ordinary differential equations* (*ODEs*). While the *SCM* of a random variable y defines such a variable as a function of their causal parents (direct causes), the *causal kinetic model* assumes the form

$$\dot{Y}_t = \frac{d}{d_t} y_t := f^y(\mathbf{x}_t^{\mathbf{PA}_y}), \tag{6.1}$$

where $\dot{Y}_t$ is the rate of change of the target variable, $\frac{d}{d_t} y_t$ denotes the derivative of y (a target variable) at time t, and $\mathbf{x}_t^{\mathbf{PA}_y}$ are the values of the set of parents of y at time t. In the proposed setting, the model of a target variable y (e.g. a biomarker gene), the dynamics of which have changed during cancer owing to the influence of driver genes, is assumed to adopt the form shown in eq. 6.1. Consequently, it is hypothesised that the set of parents of y (namely $\mathbf{PA}_y$) are cancer drivers.

6.2.1.2 Invariance of structure of dynamic systems for causal discovery of cancer driver genes

One of the most notable properties of a causal system is that over the time that the causal structure holds, causal relationships remain *invariant* even under different interventions and/or environments [13, 249]. This *invariance* property is reflected, for example, in the fact that the conditional distribution of a variable y given its parents $\mathbf{PA}_y$ remains the same under interventions on variables other than y. One of the benefits of the *invariance* property is that it can be explored to assess causality.

In the proposed method, given a target variable y with dynamics that are influenced by cancer drivers, it is assumed that observed values of y have been generated by a causal model that is *invariant* during cancer progression. As the dynamics of y is influenced by cancer drivers, the derivative of y can be modelled as a function of its causal parents (i.e. $f^y(\mathbf{x}_t^{\mathbf{PA}_y})$, as shown in eq. (6.1)). Therefore, the fitted values resulting from the evaluation of f^y will be reasonable estimations of the rate of change $\dot{Y}_t$ (from data) throughout the progression of the cancer when f^y is properly defined. This feature enables the dynamics of y to be used for identifying its causal parents.

Because gene expression data in cancer corresponds to real observations of the disease, it is possible to infer a function that describes the rate of change of a target variable y when a temporal dimension is included in the data. The *invariant* aspect of the model can be assessed by exploring cancer data collected from different "environments" (e.g. two repetitions of an experiment in which data on the same type of cancer are collected). As such data are uncommon in practice, the proposed method has the ability to emulate two different environments by systematically splitting the data (the details are in section 6.2.4). The data on the two environments are used to verify whether the model fits the values of the derivative estimated from the data on the target variable in both environments.

6.2.2 Method overview

The proposed method utilises cancer gene expression and the concept of *pseudotime* to emulate two environments and to create one pseudotemporally ordered data matrix for each environment. Subsequently, models of the form shown in Eq. (6.1) (the candidate causal structures) containing different combinations of putative drivers of the target y are tested in terms of their *invariance* across the environments. The better a model explains the data in both environments, the better the ranking towards *invariance* is given to that model. In other words, if a model provides consistent estimations of the derivative of y in both environments, then the proposed method ranks the model in a high position.

The outcome of the proposed method is a list of the top-ranked models as well as a list of cancer driver genes ranked in terms of their importance for those *invariant* models. The more times a gene appears in a highly ranked *invariant* model, the higher its rank as a cancer driver gene.

As shown in Fig. 6.1, the proposed method contains the following three phases:

1. **Formulation of putative models**: A set of models (candidate causal structures) that describe the rate of change of the target variable y as a function of putative cancer drivers is created. Only genes with a reasonable potential of being cancer drivers (see section 6.2.3 for details) are kept in the putative models. These genes are selected by using differentially expressed gene (DEG) analysis and a protein-protein interaction (PPI) analysis.

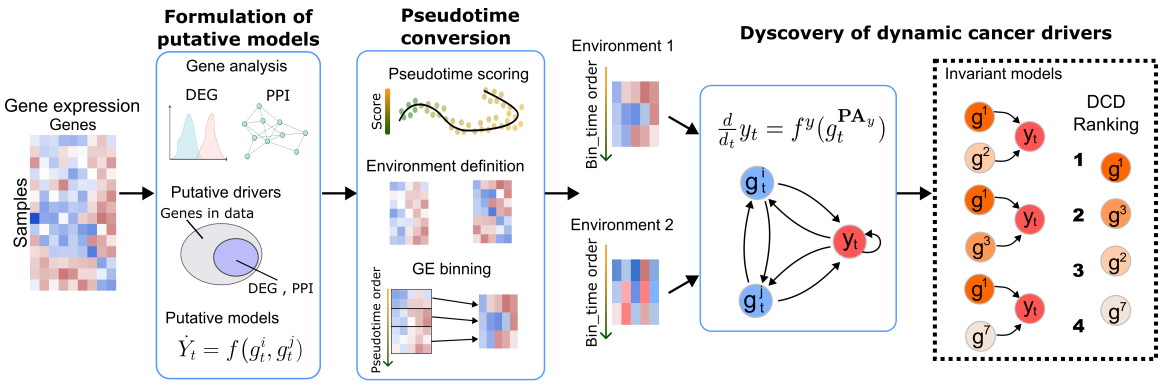


Figure 6.1: **Summary of the proposed method for identifying cooperative genes that cause cancer progression.** Starting with a gene expression dataset, the proposed method performs a putative model definition to create models that include only genes that have been selected as candidate cancer drivers by the method. Genes selected as putative cancer drivers are used for inferring a pseudotime order of samples that encode significant aspects of cancer progression. Ordered samples are grouped to form two mutually exclusive sub-datasets to emulate two different experimental environments. The gene expression and pseudotime score are binned to obtain Bin_time series containing their expected values in each environment. Bin_time series and the concept of invariance are used for finding the putative models that better describe the rate of change of a target variable y in both environments. Models that better explain the rate of change of y are considered invariant models. Genes in the invariant models are considered dynamic cancer drivers.

2. **Pseudotime conversion**: Genes selected in the first phase are utilised for *pseudotime* scoring so that a temporal dimension is created and added into the data. Both the gene expression and *pseudotime* score are binned to emulate two environments and a single *pseudotime* (named *Bin_time*).

3. **Discovery of dynamic cancer drivers**: A set of models are tested to observe how well they fit in both environments. The best fitted models are considered *stable towards invariance*, and genes consistently appearing in such models are selected as cancer driver genes (more details are in section 6.2.5). The results of this phase are a list of the dynamic models that are the most *stable towards invariance* as well as the set of cancer driver genes discovered in the data.

The details of the three phases are explained in the following sections.

6.2.3 Formulation of putative models

Gene expression is the result of a complex series of chemical reactions that can be modelled as random variables. Consequently, the *law of the mass action* (an essential

principle in chemical reactions) is a reasonable selection of a model that fits the form of in Eq. (6.1). This phase refers to the formulation of putative models (i.e. symbolic notation). Parameter estimations and model evaluations are performed in later phases.

In the proposed method, the parametric version of the *law of the mass action* [249] is adapted to the setup of the problem:

$$\begin{aligned} \dot{Y}_{\hat{t}} = \theta_{0,i} \cdot g^i_{\hat{t}} + \theta_{0,j} \cdot g^j_{\hat{t}} + \theta_{i,j} \cdot g^i_{\hat{t}} \cdot g^j_{\hat{t}}, \\ \forall\ i, j \in \{1 : m\} \text{ and } i \neq j. \end{aligned} \tag{6.2}$$

Specifically, models described by Eq. (6.2) correspond to the pairwise *main effect models* (as defined in [249]) of any pair of genes g^i and g^j. In a *main effect model*, if the interaction term exists (i.e. $\theta_{i,j} \neq 0$), then the main effect terms also exist (i.e. $\theta_{0,i} \neq 0, \theta_{0,j} \neq 0$).

In addition to the pairwise *main effect models*, two lower order family models derived from Eq. (6.2) are included in the analyses: the pairwise gene collaboration linear models (i.e. $\theta_{0,i} \neq 0, \theta_{0,j} \neq 0$, and $\theta_{i,j} = 0$) and the pairwise interaction models (i.e. $\theta_{0,i} = 0, \theta_{0,j} = 0$, and $\theta_{i,j} \neq 0$).

In Eq. (6.2), $\dot{Y}_{\hat{t}}$ denotes the rate of change of the target variable, $g^i_{\hat{t}}$ and $g^j_{\hat{t}}$ stand for the gene expression of genes g^i and g^j at $\hat{t}$, respectively, $g^i_{\hat{t}} \cdot g^j_{\hat{t}}$ represents the interaction between g^i and g^j at $\hat{t}$, and $\theta_{0,i}$, $\theta_{0,j}$, and $\theta_{I,j}$ are model parameters. In this notation, $\hat{t}$ is explicitly used (instead of t) to highlight the fact that the parameter estimation does not use real time but a pseudotemporal reference (see section 6.2.4).

The analysis of gene interactions in data falls under the category of large scale system analysis. Ideally, a comparison among all possible combinations of different genes should be performed. However, a full exploration of the search space is infeasible due to the combinatorial nature of the problem.

To better illustrate this problem, a setup with only three genes of interest $\{g_1, g_2, g_3\}$ is assumed. Specifically in this setup, nine models are tested: the three pairwise *main effect models* (i.e. models with parameters $[\theta_{0,1}, \theta_{0,2}, \theta_{1,2}]$,$[\theta_{0,1}, \theta_{0,3}, \theta_{1,3}]$, and $[\theta_{0,2}, \theta_{0,3}, \theta_{2,3}]$), the three pairwise linear models (i.e. models with parameters

$[\theta_{0,1}, \theta_{0,2}]$, $[\theta_{0,1}, \theta_{0,3}]$, and $[\theta_{0,1}, \theta_{0,3}]$), and the three pairwise interaction models (i.e. models with parameters $[\theta_{1,2}]$, $[\theta_{1,3}]$, $[\theta_{2,3}]$). Following a similar analysis, the search space expands to 18 models when considering four genes (as there are six pairwise *main effect models* in that scenario).

In general, there are $\frac{d!}{2\cdot(d-2)!}$ *main effect models* with the form shown in Eq. (6.2). As a result, the search space to be explored by the proposed method (that also includes the two lower-order family models) would consist of $\frac{3\cdot d!}{2\cdot(d-2)!}$ non-null models. This easily becomes infeasible for a large number of genes. Furthermore, models representing real driver gene interactions are sparse within the search space, as a sparsity of such interactions is expected. Accordingly, in this phase the number of models to be explored is reduced by including only genes that have a reasonable potential of being cancer driver genes.

In the proposed method, a gene has reasonable potential to be a cancer driver if it fulfills the following requirements:

1. The gene is differentially expressed between two meaningful conditions (such as normal vs cancer) in the dataset. This requirement is considered as fulfilled if such a gene is detected as differentially expressed after a DEG analysis.

2. The gene has the "potential" of influencing other genes. This requirement is considered as fulfilled if such a gene belongs to the PPI network. A cutoff of the degree of the gene (understood as the number of its interactions in the PPI network) can be used to further reduce the number of predictors when needed. The larger the degree, the more influential the gene.

Satisfying these requirements leads to a dimensionality reduction for the gene expression matrix from $\mathbf{G} \in \mathbb{R}^{n\times m}$ to $\mathbf{G} \in \mathbb{R}^{n\times d}$, where d is the number of putative drivers. Using the d putative drivers as predictors, the family of models to be analysed by the proposed method is defined as shown in Eq. (6.2).

These considerations entail a significant reduction in the number of models to be explored while maintaining the predictive power of the proposed method.

6.2.4 Pseudotime conversion

This phase uses as input the dimensionally reduced gene expression matrix $\mathbf{G} \in \mathbb{R}^{n \times d}$. This gene expression matrix is used for finding a *pseudotime* scoring that reflects the relative position of the samples throughout cancer progression. In the experiments, *Phenopath* [84] was used to calculate such a *pseudotime* score. *Phenopath* was selected because it is (to the best of our knowledge) the only *pseudotime* method explicitly designed for both single-cell and bulk data.

Inferred *pseudotime* is used for the emulation of two environments (this is required in the proposed method for testing the *invariance* property). The two environments are emulated as follows:

1. Use $\mathbf{G}$ to find a *pseudotime* from the data.

2. Reorder the samples in $\mathbf{G}$ to follow that *pseudotime* in ascending order.

3. Split the ordered dataset by keeping the samples indexed by odd numbers in environment 1 and the samples indexed by even numbers in environment 2.

4. Split the *pseudotime* vector (ordered in ascending order) by keeping the values indexed by odd numbers as the *pseudotime* for environment 1 and the samples indexed by even numbers as the *pseudotime* for environment 2.

The dataset is split using this heuristic rather than random sampling in order to conserve the samples reflecting a similar progression of the disease in both environments.

At this stage, each environment has one half of the original samples and its corresponding *pseudotime* vector, with both following *pseudotime* ascending order. In the next step in this phase, the proposed method calculates the "binned version" of both the gene expression and *pseudotime* in each environment. To find the "binned version" of a vector (including the *pseudotime* vector and the gene vectors) in one environment, the following steps are performed:

1. Split the vector using the equal frequency method with a user-defined number of bins.

2. Map each bin into its expected value (i.e. mean).

This process is applied to all genes and *pseudotime* vectors in both environments. The result of this step consists of two matrices containing the "binned version" of genes, as well as the "binned version" of each environment's *pseudotime*. Such matrices resemble the temporal progression of the disease because "binned versions" are created from samples ordered by following the *pseudotime*.

Finally, a common temporal reference for both environments is created by calculating the element-wise average of both *pseudotime* vectors. Hereafter, this common temporal reference is called *Bin_time* (notated as $\hat{t}$ in Eq. (6.2)). As a result of this process, each environment has its own *Bin_time* series for the gene expression of each gene and the target, with a common temporal reference (i.e. the *Bin_time* vector).

6.2.5 Discovery of dynamic cancer drivers

The *Bin_time* series obtained from the *pseudotime* conversion phase are used for assessing the stability of the models described in Eq. (6.2) in terms of their *invariance* across both environments. Testing whether a model is stable towards *invariance* can be achieved by testing if the estimations of the derivative of the target variable $y_{\hat{t}}$ calculated from the observed values of y can be adequately fitted by models with the form shown in Eq. (6.2) [251].

Such stability towards *invariance* can be assessed by the *stability score* of the model [251]. The *stability score* is a goodness-of-fit metric that uses the residual sum of the squares of a fitted spline as an indicator of *non-invariance*. Moreover, this score can further guide a ranking procedure for the predictors (a detailed proof of the validity of both can be found in [249]). In the proposed method, the following steps are performed to calculate the *stability score* of a model M that has the form of the Eq. (6.2):

1. Fit a solution to the model M by using the data for the target and the predictors from environment 2. The *integrated model* described in [249] is replicated.

This model implements the *trapezoidal rule* (a popular technique for approximating definite integrals) to obtain a numerical approximation of the model M as a function of the integral of its predictors. This approach is usually preferred over direct estimations of the derivative, as integration is numerically more stable than differentiation [252, 251]

2. Use the model parameters obtained from environment 2 to estimate the values of the derivative of the target in environment 1 by using the data for the predictors in the same environment (i.e. environment 1).

3. Fit a spline by using the data for the target in environment 1 as input and the estimated values of the target's derivative (step 2) as constraints. The function *constrained.smoothspline* from the *CausalKinetiX* package [251] was used to fit such a spline. The residuals of the spline are expected to be small when the parameters found from environment 2 allow for an adequate fit of the spline for environment 1.

This procedure is performed a second time, but this time the order of the environments are interchanged. As a result, the proposed method creates two splines, one for environment 1 with the derivative constraint from environment 2, and one for environment 2 with the derivative constraint from environment 1. The mean of the residual sum of the squares of such splines is used as the *stability score*. The better the fit of the model in both environments, the lower the score. Consequently, in the proposed approach, models with low scores are considered to have a more stable fit towards *invariance*. Using this score leads to a ranking in which models with lower scores are better ranked. Such an inferred ranking is used for identifying cancer driver genes.

In practice, only a small fraction of the models in the search space are expected to be *invariant*. For this reason, it is suggested to consider as a starting point the top $l = d$ models (where d is the number of putative drivers) as stable models towards *invariance*. Using this cutoff, a gene g^i can be scored to reflect the proportion of stable models that depend on such a gene, as described by

$$g^i_{score} = \frac{\sum_{k=1}^{l} V^i_k}{l}. \tag{6.3}$$

In particular, one can create the binary flag $V_k^i \in \{0, 1\}$ to specify whether or not the model M_k depends on the gene g^i ($V_k^i = 1$) ($V_k^i = 0$). This score is used as an indicator of the causal impact that a gene has on cancer. Lowering the cutoff l makes the scoring procedure more strict (as the number of models considered decreases, the number of genes that appear in at least one invariant model also decreases). This potentially improves the performance of the method for certain datasets at the cost of an increase in the number of ties in a driver's scores. Genes that consistently appear in the *invariant* models have higher scores and are considered cancer drivers by the proposed method.

The outcome of the proposed method includes a list of inferred cancer drivers ranked by their contribution toward the stability of the *invariant* models as well as the structure of the top l most stable invariant models. These invariant models contain (in general) interactions (interaction terms) and collaborations (main effect terms) of genes, as described by Eq. (6.2). Using this information, the proposed method creates a *gene collaborative network* by using any two genes in the invariant models as nodes and creating a link between genes that appear simultaneously in the same invariant model.

6.3 Experiments and results

6.3.1 Cancer drivers discovered from TCGA breast cancer data

Bulk data account for the vast majority of cancer data. Hence, experiments using the TCGA-BRCA dataset [69] were performed. The dataset was accessed in June 2022 and it was downloaded by using the *TCGAbiolinks* package in R [237]. The dataset used in the experiments contained the gene expressions harmonised against GRCh38 (hg38) of 23694 genes from 1219 samples (113 normal tissue and 1106 primary tumour). As part of the requirements of the proposed method, a target variable y with dynamics that have been altered during cancer needs to be provided. Well-known biomarkers or any gene with a gene expression that is altered because of cancer are suitable target variables provided that such alterations are captured in the analysed dataset.

In the experiments, the 40 genes identified by [253] (see Table 6.1) were explored as possible target variables. These genes have been identified as influential in breast

cancer based on data from 2433 primary breast cancer samples. A DEG analysis and a PPI analysis were performed to select the potential cancer drivers (i.e. the predictor genes in the models). For these experiments, a DEG analysis between normal and cancerous conditions was performed by using the function *TCGAanalyze_DEA* from the *TCGABiolinks* package (fdr.cut = 0.01, logFC.cut = 1, pipeline = "edgeR", and method = "glmLRT").

Table 6.1: **List of the 40 Mut-driver genes used in the experiments as target variables**, from [253].

SMAD4	KDM6A	CDH1	AKT1	ZFP36L1
USP9X	BRCA1	CDKN1B	TBX3	GATA3
FOXP1	TP53	PIK3CA	CHEK2	MAP3K1
MEN1	ARID1A	KMT2C	CBFB	GPS2
MLLT4	NF1	AGTR2	NCOR1	CTCF
TBL1XR1	SF3B1	FOXO3	KRAS	CTNNA1
ERBB2	MAP2K4	PTEN	CDKN2A	BAP1
PIK3R1	RB1	BRCA2	RUNX1	PBRM1

To perform the PPI analysis, the PPI network provided in [222] was used. To reduce the computational burden, a minimum of five interactions in the PPI network were selected as the threshold. This means that the putative cancer drivers are the genes identified as DEGs with more than four interactions in the PPI network. By following this procedure, 732 predictors (i.e. potential cancer drivers) were obtained. In real dynamic systems a variable can depend on itself (e.g. self-regulation); similarly, in the experiments the target gene y was included as an additional predictor variable. Consequently, 733 variables (i.e. one target + 732 predictors) and a search space of 804834 models (as in Eq. (6.2)) per target gene were explored. Only 39 out of the 40 possible targets were further explored because the expression of the gene AGTR2 (ENSG00000180772) was missing from the dataset.

Pseudotime conversion was performed by using the *pseudotime* score inferred by *Phenopath* [84]. The 732 genes selected in the previous phase were used for *pseudotime* scoring. *Phenopath* requires a "path covariate" as additional input to guide the *pseudotime* scoring process so that it can be used in bulk data. The gene expression of ERBB2 (HER2) was selected to fulfill *Phenopath's* requirement because this gene is both a well-known breast cancer driver and breast cancer biomarker frequently used for diagnosis and prognosis. The binning process was performed by placing 12 consecutive samples into one bin ($\approx$ 1% of the samples of the dataset) after

the creation of the two environments. As a result, *Bin_time* series with 50 *Bin_time* points were obtained for each environment.

6.3.1.1 Systematic detection of cancer drivers by the proposed method

Assessing the performance of methods for discovering cancer drivers is a difficult task because no gold standard is available, and thus the "false positive rate" is not indicative of the performance of a method. However, it is a common practice to assess a method's ability to systematically detect driver genes by using the Cancer Gene Census (CGC) [220] as a conservative approximation of the ground truth. The underlying logic of this common approach is that a method identifying more CGC genes from data than that expected by random selection (significance level ≤ 0.05) is less prone to misidentifications [254]. In the experiments, the hypergeometric test was used to evaluate whether CGC genes (CGC v94 n = 719 genes) can be systematically identified by using the driver score (Eq. (6.3)) inferred by the proposed method.

Table 6.2: **Summary of the 11 target genes with a significant overlap with the CGC gene set.** In the proposed method, the hypergeometric p.value is used as an indicator of systematic identification of cancer driver genes. The p.values were calculated by using the *phyper* function in R.

Target y	CGC baseline	n inferred	CGC detected	p.value
MEN1	94	351	56	0.0101
AFDN	94	356	54	0.0413
PIK3R1	93	445	69	0.0027
TP53	94	277	44	0.0355
NF1	94	317	50	0.0246
PIK3CA	94	351	53	0.0488
FOXO3	94	367	56	0.0309
BRCA2	93	327	53	0.0073
CHEK2	94	278	45	0.0228
CBFB	94	282	44	0.0488
CDKN2A	93	198	38	0.0013

The hypergeometric test p.value was calculated by using the *phyper* function in R (p.value = 1 - *phyper(q - 1,m,n,k)*). The parameters of the test were as follows: q = number of CGC genes among the inferred cancer driver genes, m = number of CGC genes among the 733 model variables (baseline), n = number of non-CGC genes among the putative drivers, and k = number of inferred cancer drivers. The

proposed method detects genes acting together (i.e. interacting and/or collaborating) to drive cancer progression rather than single gene drivers. Consequently, the hypergeometric test is used as an indicator of the systematic performance of the method as a whole, not as a performance metric of the method's ability to detect single drivers nor of single experiments (i.e. a single target variable).

The inferences of the proposed method have a statistically significant overlap with the CGC gene set (p.value ≤ 0.05) when using the 11 target genes shown in Table 6.2. Furthermore, a significant overlap with the CGC set was found for seven additional targets when using the driver score (Eq. (6.3)) as the metric to select the top 150 inferred genes (see Table 6.3). These results, as well as the fact that the inferred sets of cancer drivers have relatively small sizes, suggest that the proposed method is able to effectively detect and rank cancer driver genes from bulk data. The results of the hypergeometric test for the 39 genes can be found in Appendix C.1. Additional results that were obtained when analysing different top lists of inferred drivers from bulk data can be found online at https://github.com/AndresMCB/DynamicCancerDriverKM in *"supp_Table1 - hypertest for 39 target genes(Bulk).csv"* (in the *supplementary* folder).

Table 6.3: **Summary of the seven additional targets with a statistically significant overlap with the CGC gene list when using the top 150 genes ranked by the proposed method.** The p.value resulting from the hypergeometric test is shown in the last column. The p.values were calculated by using the *phyper* function in R.

Target y	CGC baseline	CGC in top 150	p.value
SMAD4	94	27	0.0261
FOXP1	94	26	0.0462
CDH1	94	30	0.0034
KMT2C	94	30	0.0034
NCOR1	94	27	0.0261
CTNNA1	93	29	0.005
PBRM1	94	27	0.0261

Besides the ability to systematically detect driver genes, an effective cancer driver discovery method is expected to retrieve genes that are strongly linked to cancer in biological terms. To assess whether the inferred cancer drivers are enriched in terms of cancer, the following analyses were performed via the *Enrichr* online tool [255]: *KEGG 2021 Human* (biological pathways), *GO Biological Processes 2021*, and *DisGeNET* (human diseases). The experiments for the targets PIK3R1 and

CDKN2A were selected for these analyses, as they have the largest (445 genes) and the smallest (198 genes) inferred sets, respectively, and significantly overlap with the CGC list.

In both cases, the inferred set was significantly enriched with respect to cancer. For both analysed targets, the top one *KEGG* enriched term (ranked by p.value) was **pathways in cancer** (PIK3R1set p.val = 3.609e-27, CDKN2Aset p.val = 1.544e-15), and all other terms in the top four were pathways related to cancer, including **focal adhesion** (PIK3R1set p.val = 1.464e-16, CDKN2Aset p.val = 1.609e-12), **transcriptional misregulation in cancer** (PIK3R1set p.val = 3.428e-16), and **proteoglycans in cancer** (CDKN2Aset p.val = 2.297e-10).

Similarly, all top five *DisGeNET* and all top five *GO Biological Processes* enriched terms (ranked by p.value) for both inferred sets were strongly related to cancer. The top five *DisGeNET* enriched terms included **neoplasm metastasis** (PIK3R1set p.val = 3.366e-63, CDKN2Aset p.val = 2.894e-50), **breast carcinoma** (PIK3R1set p.val = 4.476e-64, CDKN2Aset p.val = 8.765e-47), and **mammary neoplasms** (PIK3R1set p.val = 3.366e-63, CDKN2Aset p.val = 3.720e-54). The top five *GO Biological Processes* enriched terms include **positive regulation of transcription, DNA-templated (GO: 0045893)** (PIK3R1set p.val = 5.884e-28, CDKN2Aset p.val = 1.213e-16), and **positive regulation of transcription by RNA polymerase II (GO:0045944)** (PIK3R1set p.val = 1.479e-24, CDKN2Aset p.val = 1.144e-15). The complete results for the enrichment analyses using bulk data can be found online at `https://github.com/AndresMCB/DynamicCancerDriverKM` in the *supplementary* folder (*supplementary tables 3-8*).

6.3.2 Identification of mutated and non-mutated cancer drivers by the proposed method

For this experiment, the single nucleotide variant (SNV) frequency was included in the set of inferred cancer drivers to assess the ability of the proposed method to identify cancer drivers regardless of their mutational status. This is one of the key aspects of the proposed method, as not all cancer driver genes are mutated.

Because non-mutated cancer drivers have been investigated minimally in the literature, it is difficult to evaluate the performance of non-mutated driver inference methods. Thus, this analysis was restricted to highly reliable driver genes inferred

by the proposed method because the genes that have been consistently detected as cancer drivers in many of the experiments were less likely to be false positives. Specifically, only the genes that were detected as cancer drivers in more than 50% of the 11 experiments summarised in Table 6.2 were considered. The SNV frequency was incorporated into the resulting set, whicht contained 308 genes. Mutational information was obtained from the TCGA-BRCA dataset in Mutation Annotation Format (MAF) (MAF.hg38) by using the R package *TCGAbiolinks*.

The SNV frequency was used to classify the cancer driver discoveries as mutated (n=280) or non-mutated (n=28). The Network of Cancer Genes & Healthy Drivers (NCG 7.0) was used to further analyse the top 10 genes with the highest SNV frequency and all non-mutated genes [256]. The NCG is a manually curated collection of cancer genes that contains canonical cancer drivers (such as the CGC genes) and highly confident candidate drivers.

Table 6.4: **Top 10 most mutated (SNV) and all non-mutated inferred cancer drivers in the intersection set of the results for the 11 target genes shown in Table 6.2**. Only genes identified in more than 50% of the 11 experiments (i.e. 6 or more) were considered. Canonical (CGC) and highly confident drivers (NCG) are shown in boldface font. The NCG7.0 analysis was performed using the online tool at `http://ncg.kcl.ac.uk/`.

Top 10 Mutated Drivers	**SYNE1**, **DST**, **DMD**, **CUBN**, SVEP1, **ANK2**, UTRN, **LRP1**, **F8**, **PTPRB**
Non mutated Drivers	**IRS2**, PECAM1, CXCL12, MDK, **CALR**, CIB1, CAV2, **ALDH2**, CLEC3B, TK1, AQP1, SCTR, ANTKMT, ENO2, FADD, LEP, PRTN3, LURAP1, SNCG, NME3, **CDKN2A**, PPP1R14A, CCL19, KCTD13, MAFF, SIGIRR, **IRF7**, **LTB**

As expected, the proposed method was able to consistently recover canonical/candidate cancer drivers, including non-mutated drivers (see Table 6.4). Seven out of the top 10 most mutated inferred genes (namely SYNE1, DST, DMD, CUBN, ANK2, LRP1, and F8) are NCG genes while PTPRB is a CGC gene with the role of a tumour suppressor. The genes CALR, ALDH2, and CDKN2A are CGC genes identified by the proposed method that have no mutation in the TCGA-BRCA dataset. Additionally, the proposed method identified three other highly confident drivers (i.e. IRS2, IRF7, LBT) that have no mutation in the analysed dataset.

To name a few of the novel non-mutated cancer drivers discovered by the proposed method, CXCL12 (specifically the CXCL12/CXCR4 signal pathway) has been found to promote several aspects of breast cancer tumorigenesis, such as proliferation, cell motility, and distant metastasis [257, 258, 259]. TK1 and MDK have also been linked to advanced/metastatic stages of breast cancer [260, 261, 262, 263, 264], while CIB1 is a promising therapeutic target gene that drives cancer cell death [265, 266]. The relevance of the discovered genes to cancer provides strong support for the effectiveness of the proposed method in identifying mutated and non-mutated drivers of cancer.

6.3.3 Inferred cancer driver sets from single-cell data

To assess the functionality of the proposed method for single-cell data, the single-cell RNA sequencing data from the NCBI GEO database (accession GSE75688) [219] was analysed. The original dataset contains gene expressions of 515 single cells from 11 breast cancer patients with distinct molecular subtypes: oestrogen receptor positive (ER+), human epidermal growth factor receptor 2 positive (HER2+), double positive (ER+ and HER2+), and triple-negative breast cancer (TNBC). Out of 515 cells in this dataset, 317 correspond to tumour cells while the other 198 correspond to immune (n=175) and stromal (n=23) cells. After removing low expressed genes, the dataset had 9288 genes.

For these experiments, the putative model formulation phase was performed by including differentially expressed genes with a node degree of one in the PPI network (i.e. genes with at least one interaction in the PPI). A DEG analysis (fdr.cut = 0.01 , logFC.cut = 1, pipeline = "edgeR", and method = "glmLRT") between the cancerous and normal conditions was performed by using the single cells labelled as "tumour" for cancer and the cells labelled as "stromal" for normal. This procedure led to the selection of 178 genes for further exploration. A total of 340 samples were used for the data-driven inference process (tumour and stromal cells). For the *pseudotime* conversion phase, the number of *Bin_time* points was set to 30 to allow at least five samples per *Bin_time* point during the binning process.

The 39 Mut-driver genes in Table 6.1 that are present in the dataset were used as possible targets (AFDN was missing). The results for all 39 genes were similar in terms of the size of the inferred sets (84-129, mean = 105, sd = 9.11) and the size of the intersection with the CGC set (12-19, mean = 14.79, sd = 1.67). The proposed

Table 6.5: **Summary of single-cell data experiments for the 14 target genes with a significant overlap with the CGC gene set when considering the top 80 inferred genes per target**. The p.values were calculated by using the *phyper* function in R.

Target y	CGC baseline	n_inferred	n_CGC	p.value (top 80)
SMAD4	22	94	14	0.0467
SF3B1	22	98	15	0.0162
RB1	22	109	14	0.0467
CDH1	22	108	14	0.0467
CDKN1B	22	97	15	0.0162
PIK3CA	21	101	14	0.0289
FOXO3	22	100	15	0.0162
PTEN	22	88	15	0.0162
AKT1	22	114	15	0.0162
TBX3	22	108	15	0.0162
CHEK2	22	111	14	0.0467
RUNX1	22	93	14	0.0467
ZFP36L1	21	110	14	0.0273
MAP3K1	22	113	17	0.0010

method's inferences showed a significant overlap with the CGC gene set for seven of the explored target genes (SMAD4, RB1, CDKN1B, FOX03, PTEN, ZFP36l1, and MAP3K1) when considering all genes with a score (as calculated by using Eq. (6.3)) greater than zero. The number of targets with a significant overlap increased to 14 (see Table 6.5) when considering the top 80 genes of each target (selected as the cutoff because the size of the smaller inferred set among the single-cell experiments was 84).

The inferred driver sets from those 14 targets were further explored to assemble a single driver list. Table 6.6 shows the cancer driver genes inferred from the single-cell data that were detected by the proposed method for at least 10 different target genes (i.e. they appear in the intersection of at least 10 of the inferred sets). The genes ATP1B1, EZR, FBLN1, HLA-E, and TAPBP were identified as cancer drivers by the proposed method in all 14 targets, suggesting these genes are highly relevant to cancer progression for the 11 patients of the study. Among those five genes, EZR is a CGC gene, while ATP1B1 and FBLN1 are designated as *mouse genes* in the *Catalogue Of Somatic Mutations In Cancer (COSMIC)* [267]. A *COSMIC mouse gene* is a gene with experimental support (mouse insertional mutagenesis) that strongly suggests such a gene as a cancer-causing gene. The full list of inferred cancer drivers from the single-cell data for each of the 39 experiments can be found

Table 6.6: **Cancer driver genes inferred from the single-cell data that were detected by the proposed method in at least 10 different experiments (different target genes).**

n targets y	Inferred drivers
14	ATP1B1, EZR, FBLN1, HLA-E, TAPBP
13	CALU, CD99, DHX9, DPYSL2, ITM2B, PDIA4, RNF19A, SERPING1, SNRPF, STOM, SYNE1, TPD52
12	ARF1, AXL, CDK12, COL1A1, FLNA, GLRX3, GSN, HLA-C, IFI16, LRPPRC, PIK3CA, SND1, UBR5, VIM, ZNF24
11	ABCA1, ACTR1B, BCCIP, C3, DAB2, ERBB2, FER, FOXP1, FTL, HSP90AB1, IGFBP4, IGFBP7, ITGB1, LAMP1, MUC1, OCLN, PRSS23, PTPRF, RIN2, RPS6KA3, SIX1, STAT1, STAT2, TNPO2, TOB1, WWTR1
10	ANXA5, B2M, DNAJC10, ERBB3, GLI3, IKZF5, JPT2, PEX13, PTGFRN, TAX1BP1, TCF4, TFAP2A, TFAP2C, TGFB3, UBA1

at https://github.com/AndresMCB/DynamicCancerDriverKM in *"supp_Table9 - inferred drivers from single cell.xlsx"* (the *supplementary* folder).

As the proposed method detects genes that act together (i.e. collaborate and/or interact) to cause cancer to progress rather than single gene drivers, the group gene cooperation networks obtained from the experiments were examined. Specifically, the networks inferred when using PTEN and AKT1 as target variables (hereafter referred to as **PTENnetSC** and **AKT1netSC**, respectively) were explored. These two experiments were chosen because they returned the smallest and largest sets of inferred cancer drivers, respectively. A community detection process was performed by using the function *group_infomap* from the *tidygraph* R package [268] and using the default parameters (see Fig. 6.2).

For the sake of visualisation, the networks were built by only considering the top 50 models per target ranked by their *stability score* (as described in section 6.2.5). The network nodes correspond to the genes appearing in any of the top *invariant* models. The edges in the networks link genes appearing simultaneously in at least one of the top *invariant* models. As expected, the inferred networks included canonical driver genes. Specifically, FLNA, B2M, H3-3A, ARID2, and FOXP1 in **AKT1netSC**, and COL1A1, ERBB3, CDK12, PIK3CA, EZR, SND1, ELF3, PTEN,

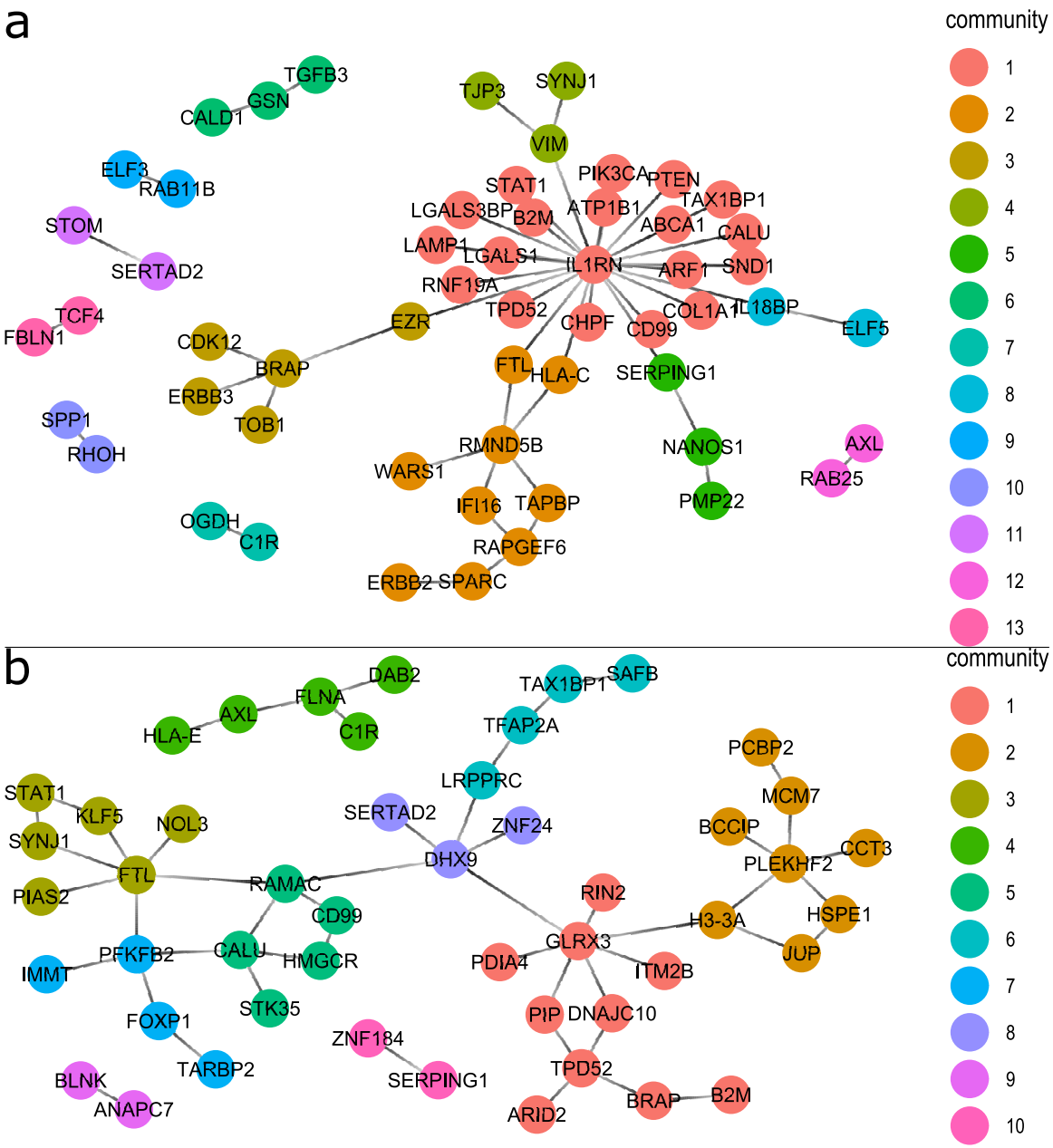


Figure 6.2: **Inferred gene cooperation networks when using the top 50 models of the experiments with a) PTEN (PTENnetSC) and b) AKT1 (AKT1netSC) as target variables.**

B2M, ERBB2, and RHOH in **PTENnetSC**.

The inferred novel drivers FTL, DHX9, GLRX3, and PLEKHF2 were central nodes in **AKT1netSC**, while IL1RN and RMND5B were central nodes in **PTENnetSC**. FTL, PLEKHF2, and DHX9 are *COSMIC mouse genes*, and DHX9 is also a putative oncogene in the NCG 7.0 database. Additionally, it has been confirmed that IL1RN is a tumour-associated gene that is negatively correlated with the survival time

of glioma patients [269], and it has also been found to be upregulated across the ephithelial-to-mesenchymal transition (EMT) in triple negative breast cancer [71].

Finally, the community surrounding the IL1RN gene (i.e. community 1 in Fig. 6.2-a) was explored, as it was the largest community detected among these two networks (19 genes). A series of pathway analyses was performed to verify the biological relevance of the inferred cluster. *Reactome 2022*, *KEGG 2021 Human*, and *Elservier Pathway collection* enrichment analyses were performed via the *Enrichr* online tool [255]. Within the top five enriched terms of each analysis (ranked by combined score), four *Reactome 2022*, one *KEGG 2021 Human*, and four *Elservier Pathway collection* enriched terms were strongly linked to cancer (see Table 6.7).

Table 6.7: **Enriched biological pathways for the IL1RN gene cluster (community 1 in Fig. 6.2a) in Reactome 2022, KEGG human 2021, and Elsevier Pathway Collection.** The IL1RN inferred gene community was enriched in several biological pathways closely related to cancer. The enrichment analysis was performed by using Enrichr https://maayanlab.cloud/Enrichr/.

Database	Term index	Enriched Terms	P.value	Combined Score
Reactome 2022	1	Signaling by PDGFRA Extracellular Domain Mutants R-HSA-9673770	0.000046	2602.87
	2	Signaling by Cytosolic FGFR1 Fusion Mutants R-HSA-1839117	0.000115	1420.01
	3	Signaling by KIT In Disease R-HSA-9669938	0.000144	1221.34
	4	Signaling by PDGFR In Disease R-HSA-9671555	0.000144	1221.34
KEGG Human 2021	2	PD-L1 expression and PD-1 checkpoint pathway in cancer	0.000078	410.07
Elsevier Pathway Collection	1	Genes with Mutations Associated with Hereditary Breast and/or Ovarian Cancer Syndrome	0.000030	3488.77
	2	Genes with Mutations Associated with Endometrial Cancer	0.000038	2987.15
	3	PI3K/AKT/MTOR Signaling Activation in Cancer	0.000115	1420.01
	5	Genes with Mutations in Cancer Metabolic Reprogramming	0.000144	1221.34

6.4 Conclusion

In this chapter, a new causal inference method for cancer driver discovery that considers the dynamics of gene interactions and collaborations to identify sets of cancer drivers was presented. The proposed method utilises the fact that driver genes alter the dynamics of core processes in cancer cells by dysregulating biological pathways. One interesting advantage of this method is that it relaxes the constraints of detecting mutational events by focusing on the observed changes of cancer driver expression and their causal influence on other genes during cancer development. This allows the proposed method to be used for identifying cancer drivers with no mutations.

The causal kinetic models of cancer drivers were considered an extension of structural causal models to the setting of ordinary differential equations, as assessing the causal effect of a driver gene on other genes to distinguish a cancer driver (the cause) from affected genes (the effects) is a sensible approach to reduce spurious discoveries. The experiments suggest that the inclusion of formal procedures for testing the causal structure of the underlying dynamics of cancer helps to infer driver genes strongly related to altered biological processes in cancer.

Another interesting aspect of the proposed method is that it produces models of gene interaction dynamics, which provide insight into the temporal behaviour of genes during cancer. Such models can be utilised to represent the interactions throughout cancer as a cooperative network. The network itself represents the set of drivers acting together (not necessarily regulatory relationships) to induce cancer progression. The inferred networks can be potentially used for identifying clusters of genes that dysregulate specific processes during cancer. Consequently, the network provides insights into the mechanisms that can cause carcinogenesis and the early development of cancer.

Chapter 7

Conclusions and future work

In this thesis, three novel methods for solving three key problems regarding the dynamics of gene regulation throughout cancer progression were presented. Specifically, the following problems were addressed: (1) the characterisation of altered biological processes in cancer, (2) the identification of genes driving the altered biological processes in cancer, and (3) the identification and modelling of the dynamics of gene interactions and collaborations between cancer driver genes during cancer development.

7.1 Summary of contributions

To solve the first problem (the characterisation of altered biological processes that occur during cancer), a method of dynamic causal inference that identifies miRNA-mRNA relationships was proposed. This method verifies the *invariance* property of the causal structures formed by miRNA-mRNA pairs throughout the entire course of the disease. The method uses high-dimensional gene expression data and incorporates a temporal dimension to the data using the concept of *pseudotime*. The inclusion of a temporal dimension in the data allows the method to search for regulatory relationships that have invariant causal structure during cancer development.

This method is available in an open-source R package called ***PTC*** (available at: https://github.com/AndresMCB/PTC). The experiments conducted using this implementation demonstrate the effectiveness of the method in terms of its ability to detect miRNA-mRNA pairs in altered biological processes that occur during cancer progression. The method outperforms widely accepted methods for detecting regulatory pairs in terms of the number of detected pairs that have been confirmed in biological experiments. These results support the utility of dynamic causal analysis in identifying regulatory interactions that define the altered biological processes that occur throughout cancer progression.

An interesting advantage of this method is that it characterises of gene regulation relationships pertaining to altered biological processes that include non-coding genes (specifically miRNAs). This is of particular interest, as it was only recently discovered that non-coding regions of DNA play an important role in the onset and development of cancer. As a result, there is limited unified information regarding the impacts that non-coding genes have on the disease. The characterisation of

regulatory relationships in cancer that include non-coding genes has been under-explored in part because of the difficulty in evaluating the functional impacts of such genes. Research such as that performed for this thesis contributes to filling the gaps in our knowledge about the non-coding regions of DNA, which comprise about 99% of the human genome.

The method we developed for solving the second problem is the freely accessible in an R package called ***DynamicCancerDrivers*** (available at: https://github.com/AndresMCB/DynamicCancerDriver). Throughout the course of the research performed for this thesis, cancer driver genes were observed to play a crucial role in advancing cancer by causing deviations from normal states to carcinogenic ones within core biological processes. The method identifies these cancer-causing genes by analysing how they causally affect biological core processes, inducing changes from their normal state to a cancerous one. The group of cancer driver genes was discovered to be neither static nor uniform; instead, it is dynamic and specific to each biological core process. In other words, different sets of genes drive cancer by influencing the dynamics of various biological processes.

One other significant contribution of this method is its ability to detect both mutated and non-mutated cancer driver genes. The standard approach for identifying cancer driver genes involves detecting positive selection signals in non-passenger mutations within putative driver genes. While this approach is valuable, it has limitations. Not only do driver genes exhibit carcinogenic aberrations that are not necessarily mutations, but this ecological (Darwinian) analysis also overlooks the dynamic nature of the regulatory relationships between driver genes. By focusing on the dynamics, the proposed method can detect cancer driver genes even if their aberrations are not mutational. This enables driver genes that are "invisible" to traditional methods to be detected.

The experiments conducted for this thesis, which included both single-cell and bulk data, demonstrated that the sets of genes inferred by the aforementioned method as drivers of different biological processes are significantly associated with cancer-related biological terms. Moreover, when the subset of genes inferred to have somatic mutations were analysed, this method outperformed state-of-the-art methods for identifying cancer drivers based on bulk data. These results suggest

that this method can complement traditional mutation-based approaches and expand the repertoire of highly reliable cancer driver genes.

Finally, to address the third problem (i.e. the dynamic modelling of genes jointly driving cancer), a method that identifies and models the dynamics of genes that collaborate and interact to drive cancer progression was developed. This method is available in an open-source R package called ***DynamicCancerDriverKM*** (available at: https://github.com/AndresMCB/DynamicCancerDriverKM). While most existing methods are designed to detect individual cancer driver genes, this method utilises the causal properties of dynamic systems to identify genes that interact and collaborate to jointly drive cancer.

One interesting feature of this method is its capacity to generate models of interactions and collaborations between cancer driver genes based on the causal analysis of the dynamic systems that such genes comprise. The method utilises these dynamic models to assemble a "collaborative" gene network. The inferred network then links the genes that appear in the same dynamic model, thus identifying them as jointly contributing to changes in cancer dynamics. This feature of the method allows for further network analysis that can potentially identify, for example, clusters of genes that dysregulate specific processes that occur during cancer.

The models and networks that were derived as a result of the development of these three methods allow the causal relationships involving cancer driver genes to be evaluated and shed light on the dynamic behaviour of these relationships throughout the course of the disease. Consequently, these methods provide valuable insights into the behaviour of cancer driver genes during cancer development, including the initial stages of the disease.

In all three methods, incorporating a temporal dimension into the gene expression data for dynamic analysis significantly enhanced the results compared to analogous methods that rely on static data. Approaching cancer through the dynamics of gene regulation allowed the critical aspects of the disease that remain imperceptible in static analyses to be visualised and better understood. Furthermore, the additional information gleaned from dynamic analyses overcame certain limitations associated with current methods. For instance, the first method can identify gene relationships, including non-coding genes. Similarly, the second and third

methods can pinpoint cancer-related genes, even when these genes lack somatic mutations.

7.2 Future directions

This thesis identified a wide spectrum of research problems for which dynamic causal analysis of gene regulations offers a promising solution. In the rest of this chapter, three potential future directions for the research presented in this thesis are presented.

1. **Development of a unified package for dynamic causal inference in cancer research**

   The development of a unified software package that integrates all three methods presented in this thesis is a natural extension of this research. Such advancements are pivotal for propelling the field of dynamic causal analysis forward, especially within the context of cancer research. Currently, the methods developed in this thesis are tailor-made for addressing three specific problems. While these methods have shown effectiveness in their respective areas, they were designed for distinct applications. Although the results of each method can complement the others, in their current state the methods lack interfaces that allow them to interact directly with each other.

   However, a significant potential exists for the integration of these methods. The similarity in the structure of the input data across all three approaches provides an opportunity to develop a transversal pipeline. Such a pipeline could facilitate the incorporation of these methods in a manner where the results from the statistical tests of one method can inform and enhance the analyses conducted by the others. Furthermore, this integration would offer a more comprehensive view of the genetic dynamics in cancer. This would be achieved by simultaneously enabling network analysis of the regulatory networks associated with altered biological processes, as well as facilitating the analysis of the dynamics of key factors involved in the onset and development of cancer.

   To date, there is no integrated software solution specifically designed for

dynamic causal analysis in the context of cancer research. A unified software package that encompasses these three methods would significantly enhance the capability to conduct dynamic causal analysis across various cancer datasets.

Another advantage of this unified software package would be its capacity to streamline the research workflow. Researchers would be able to seamlessly transition between different methods without needing to preprocess data, convert formats, or perform complex data manipulations. This streamlined workflow would not only save valuable time but also reduce the risk of errors, allowing researchers to focus on the scientific aspects of their work rather than dealing with technical challenges.

In addition to improving workflow efficiency, unifying the methods offers other potential advantages, such as the facilitation of methodological consistency (the cornerstone of scientific rigor). Furthermore, by standardising the integrated utilisation of the dynamic causal inference methods, their overall scalability and adaptability will be enhanced, which will ensure that they can be easily adapted to future developments in cancer research, be able to incorporate additional data sources, and facilitate the implementation of these frameworks in emerging research.

2. **Development of cancer prognostic methods based on causal dynamics**

The prognosis of individual cancer patients remains challenging because of the heterogeneity of the disease and the speculative nature of the problem. Currently, it is not possible to reliably verify the accuracy of prognostic methods for patients receiving or about to receive cancer treatment. In general, providing accurate prognoses for individual patients is a challenging task because of factors such as biological variability, data incompleteness, and the dynamic nature of cancer.

Recent research has shown that the concept of the "invariance" property of causal structures can be applied to the context of transfer learning in machine learning models [270, 271]. The fundamental idea in transfer learning is that when two prediction tasks share common elements (i.e. they are similar), a

model trained on one task can contribute to learning a model for the second task. In this way, knowledge acquired in one model can be "transferred" to another. In the case of two domains or prediction problems with a common causal mechanism, the *causal transfer learning approach* aims to transfer a set of predictors that remain *invariant* across the domains.

While it may initially appear to be a problem of generalisation, causal transfer learning can potentially be utilised to identify the invariant characteristics of cancer from patients with known prognoses that can be transferred to enhance prognostic models. These characteristics can then serve as a basis for training tailor-made prognostic models for (similar) patients prior to treatment. Without transfer learning, such training often becomes unfeasible due to the fact that data from a single patient is usually statistically insufficient.

Furthermore, the concept of incorporating dynamic causal mechanisms to transfer learning is promising in this context as time plays a critical role in cancer prognosis. For instance, in scenarios with patients who have the same type or subtype of cancer, the problem domains are similar (the cancer type is the same) but not identical (the patients are different). In principle, it should be feasible to train models on a patient's (or population's) data and transfer valuable knowledge to models for new patients. Furthermore, integrating aspects of the causal dynamics of cancer could potentially address inaccuracies in the predictions made by current methods (which are caused by the dynamic nature of the disease).

3. **Controllability of cancer at the genetic level**

Despite the close relationship that exists between causality and control theory in dynamic systems, the areas of research and application are usually independent. This is even more pronounced in biological cancer research. On the one hand, causal theory is usually strongly biased towards static statistical analyses, and thus often neglects the dynamic aspect of systems. On the other hand, control theory, especially in the engineering field, simplifies the stochastic aspects that occur in real systems.

Although there is a trend toward the creation of intelligent control systems,

which are fundamentally statistical, the vast majority of engineering control systems in the real world do not require complex stochastic analysis. Instead, they are typically based on proportional-integral-derivative (PID) control models or state-space control systems, which are usually deterministic. Despite the proven practical usefulness of control theory over decades, its application in the biological field remains an open problem, particularly due to the lack of suitable biological models for control theory.

Recent studies have used control theory to identify critical nodes in gene regulatory networks in cancer [75, 272]. Although promising, these approaches apply a basic version of controllability concepts inherited from control theory to static representations of cancer. Therefore, they completely overlook the dynamic aspect of both control theory and cancer progression.

An extension of the models obtained in the third method (which models cancer as a collection of gene pairwise causal dynamical systems), has new potential applications when combined with control theory. According to modern control theory, if a system is dynamic there is a small subset of system variables that allow its dynamic state to be represented at any moment in time. These system variables are known in control theory as *state variables*, and modelling the system using them is known as *state-space representation*. Assuming that cancer (as a dynamic system) is controllable, then it may be possible to "control" the progression of cancer if a model of the disease in which the genes that behave as state variables are identified can be obtained.

4. **Dynamic causal modelling of spatial transcriptomics data**

This thesis has demonstrated the benefits of incorporating a temporal dimension into cancer gene expression data. Nevertheless, the detrimental effects of cancer progression are manifested not only over time but also in space. Advanced stages of many types of cancer are characterised by the presence of large malignant tumours. While metastasis can occur in the early stages of the disease, if the primary tumour has grown sufficiently large to invade surrounding tissues and release cancer cells into the bloodstream or lymphatic vessels, the expectations for the patient's survival are very low in the majority of cases.

Spatial transcriptomics is a novel technique that introduces a spatial dimension into gene expression data. This innovative data source enables the analysis of spatial patterns of gene expression within tissues. Simultaneously incorporating both temporal and spatial dimensions in gene expression data offers a promising technique in cancer research. In theory, it would enable the identification of the spatiotemporal patterns that lead to the transition of a tumour into its metastatic state. Consequently, causal models of these patterns would allow the identification of the potential factors that contribute to tumour metastasis. Understanding the spatiotemporal gene regulation patterns that drive the most aggressive states of cancer opens up new opportunities for cancer treatment.

Overall, numerous fields within biology and bioinformatics stand to gain from the exploration of novel approaches to dynamic causal inference. This chapter explored three potential future avenues for investigating causal dynamics in the context of cancer research. Furthermore, research extending beyond the realm of cancer biology also presents viable prospects for further advancements in the field.

Appendix A

Supplementary material of the PTC method

A.1 Single cell data processing

The original single cell data set [77] was downloaded from Gene Expresson Omnibus (GEO) website, under series GSE114071. The downloaded series data contains Reads Per Kilobase Million (RPKM) values for mRNAs and log2 of the normalized reads that reflect the fraction among all miRNA reads for miRNAs. Since this dataset does not contain explicitly miRNA gene expression, we applied the transformations described below:

- **Reverse log2**: Log2-transformation was reverted by calculating 2 raised to the miR power (2^{miR}). Where miR corresponds to each value provided in the original dataset. Because of the way that data was processed by [77], the reverted values represent the fraction of a given miRNA within all miRNAs (refered as miRNA expression or miRNA levels in [77]).

- **Amplification**: Reverted data was amplified by a factor of 30.18, which is the mean of total reads that map to miRNAs for all cells reported in the original study. Duplicated miRNA names were removed by selecting each first coincidence found in GSE114071.

A.2 ”GO biological processes” and ”KEGG pathways” analysis

Outcome of the ”GO biological processes” and ”KEGG pathways” analyses for the miRNAs and genes in interactions discovered by *PTC* from the processed single cell data set [77]. These analyses were performed by using the *Enrichr* web tool [205].

Table A.1: **PTC inferred genes significantly associated with KEGG pathways and GO biological processes.** A large number of genes predicted by or method are significally related with enriched terms of EMT processes. p.val = 0.00 means p.val < 0.01, Overlap column is defined as the number of inferred genes that are in the term/all genes of that term.

PTC inferred genes that are significantly associated with KEGG pathways

Term	Overlap	p.val	Genes
p53 signaling pathway	7/72	0.00	CCND3, SIVA1, GADD45A, CDK4, EI24, CDK1, CYCS
MicroRNAs in cancer	13/299	0.00	MIR10A, PRKCB, MIR25, MIR27B, MIR27A, MIR200C, RHOA, MIR125A, MIR423, DDIT4, MAPK1, VIM, MIR30D
Non-small cell lung cancer	4/66	0.02	PRKCB, GADD45A, CDK4, MAPK1
Bacterial invasion of epithelial cells	4/74	0.03	ARPC2, ARPC5, RHOA, ACTG1
Leukocyte transendothelial migration	5/112	0.04	RAP1B, PRKCB, MYL12A, RHOA, ACTG1

PTC inferred genes that are significantly related with GO Biological processes

Term	Overlap	p.val	Genes
Regulation of hematopoietic stem cell differentiation (GO:1902036)	10/76	0.00	PSMD7, PSMA4, PSMB5, YTHDF2, PSMA1, PSMD14, UBB, UBC, H3F3A, PSMB9
Regulation of hematopoietic progenitor cell differentiation (GO:1901532)	10/77	0.00	PSMD7, PSMA4, PSMB5, YTHDF2, PSMA1, PSMD14, UBB, UBC, H3F3A, PSMB9
Regulation of stem cell differentiation (GO:2000736)	10/88	0.00	PSMD7, PSMA4, PSMB5, YTHDF2, PSMA1, PSMD14, UBB, UBC, H3F3A, PSMB9
Negative regulation of cell migration involved in sprouting angiogenesis (GO:0090051)	2/14	0.02	PDCD10, RHOA
Regulation of cell division (GO:0051302)	4/65	0.02	PRC1, TXNIP, PIN1, CALM2
Regulation of cell differentiation (GO:0045595)	5/121	0.05	DDX5, PPP2R1A, PRKCB, PA2G4, ANP32B

Appendix B

Supplementary material of the Dynamic Cancer Drivers method

B.1 GO biological processes analysis

We performed a GO biological processes enrichment analysis to the single cell RNA sequencing data from NCBI GEO database, accession GSE75688 [219] to verify how significantly related are our discoveries to processes in cancer. Our analyses show a significant number of discovered drivers are strongly related relevant biological processes in cancer disease. Top ten enriched terms (for each trajectory) from Go Biological Processes are shown in Fig. B.1 . For both dynamic drivers inferred sets (i.e from VIMtime(SC), and HER2time(SC)), top 10 enriched terms (ranked by p-value) are relevant to cancer (e.g. regulation of transcription and regulation of apoptotic process).

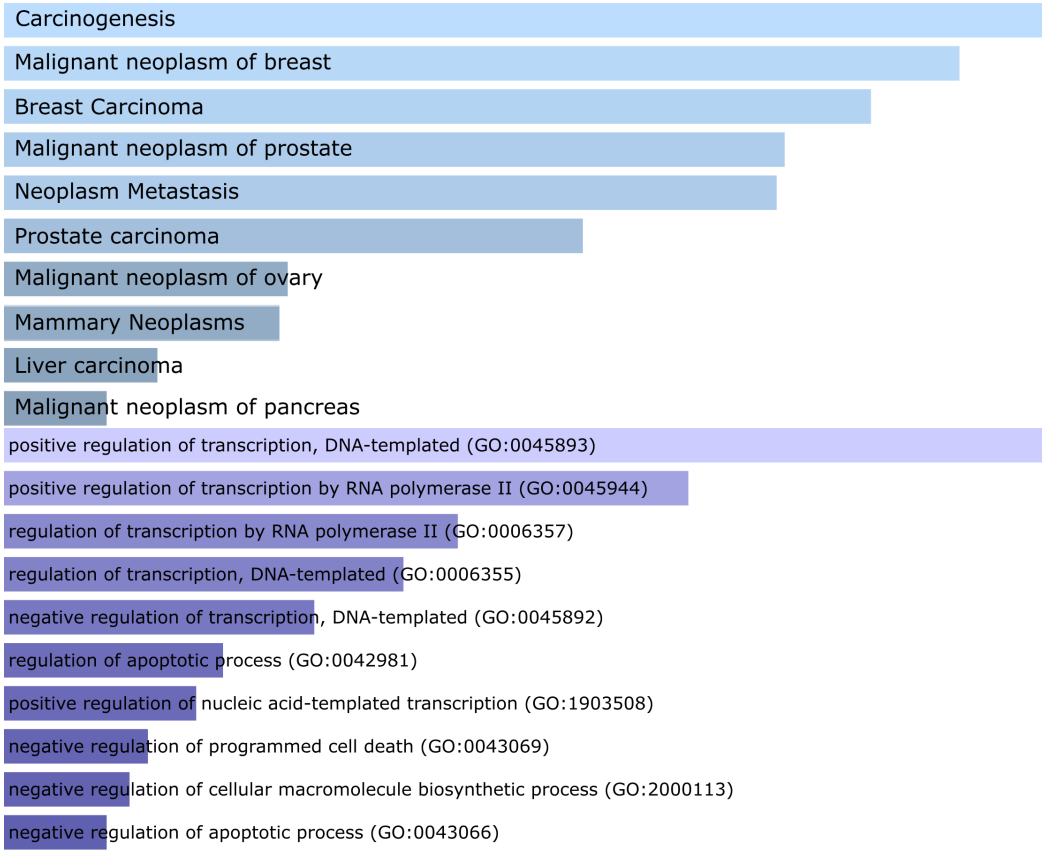


Figure B.1: Top 10 Go biological terms 2021 (ranked by p-value) from our enrichment analysis for the gene set obtained from "VIMtime(SC)" (upper) and "HER2time(SC)" (lower). In both cases, enriched terms correspond to regulation of biological processes relevant to cancer progression. Bar length represents the significance of the term. Brightness is used as auxiliary visual for significance, the lower the p-value, the brighter the colour. Analysis performed by using Enrichr [273]

B.2 Gene-disease association analysis (DisGeNet)

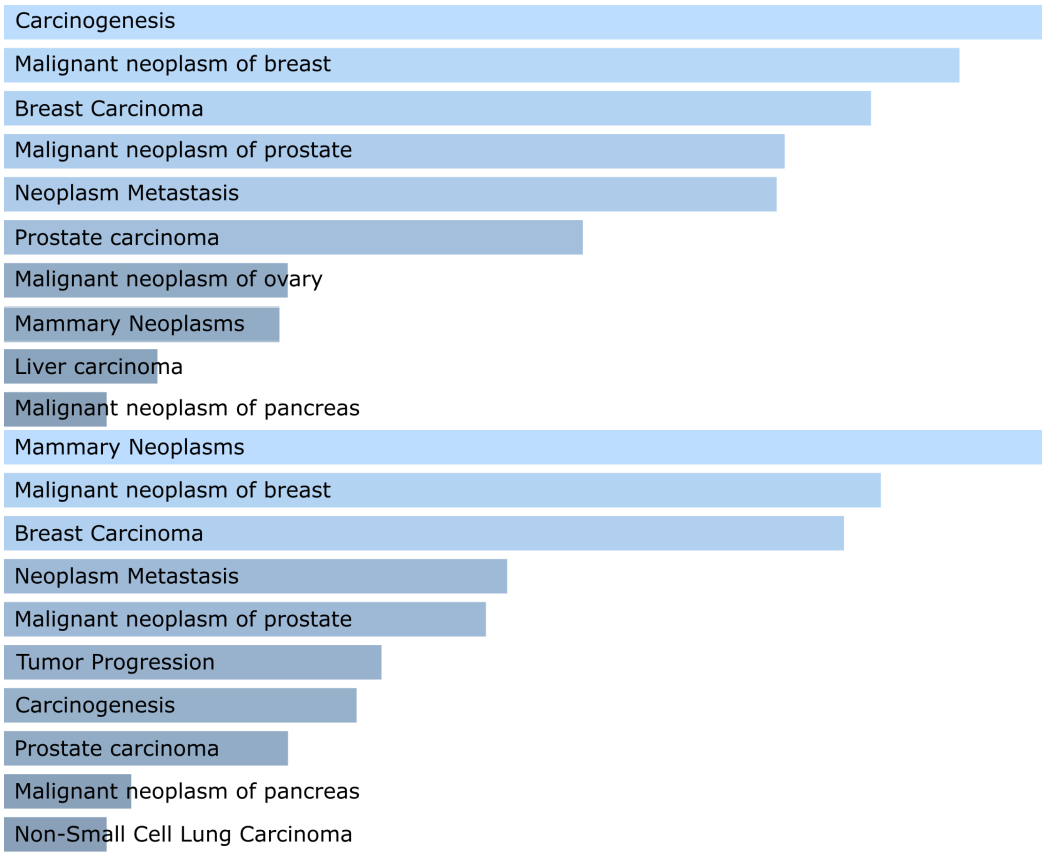


Figure B.2: Top 10 GDAs terms (ranked by p-value) from our enrichment analysis for the gene set obtained from "VIMtime(SC)" (upper) and "HER2time(SC)" (lower) retrieved from DisGeNET. Bar length represents the significance of the term. Brightness is used as auxiliary visual for significance, the lower the p-value, the brighter the colour. Analysis performed by using Enrichr [273]

We performed an enrichment analysis to identify gene-disease associations (GDAs) in the inferred drivers from the GSE75688 single cell dataset [219]. We use DisGeNET [236] as gene-disease associations database for this analysis. Our analyses show a significant number of discovered drivers are strongly related to cancer disease. The top ten enriched GDAs terms (ranked by p-value) for each of the analysed pseudotimes (i.e VIMtime(SC), and HER2time(SC)) are shown in Fig. B.2. In both cases, all of top ten terms are related to cancer disease and cancer progression.

B.3 Top 100 dynamic cancer drivers from HER2time(SC)

Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact (absolute value - descending order). A full comprehensive list with the 604 drivers inferred from HER2time(SC) can be found online in *"supplementary table 1 - dynamic cancer drivers HER2time(SC).csv"* (folder: *inst/Supplementary/*) at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.1: **Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact (absolute value - descending order).**

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
1	ENSG00000157110	RBPMS	4178.1113	18.84192	0.005
2	ENSG00000175899	A2M	1774.7859	131.9763	0.005
3	ENSG00000182568	SATB1	1367.9053	10.40657	0.005
4	ENSG00000091831	ESR1	995.47129	51.01265	0.005
5	ENSG00000108395	TRIM37	591.49912	12.86022	0.005
6	ENSG00000113263	ITK	550.15856	6.631933	0.005
7	ENSG00000163629	PTPN13	424.26875	9.785457	0.005
8	ENSG00000118785	SPP1	404.61913	62.07385	0.005
9	ENSG00000108821	COL1A1	251.6411	2.975795	0.005
10	ENSG00000163703	CRELD1	236.81685	10.45121	0.005236
11	ENSG00000186416	NKRF	178.36777	8.43429	0.005236
12	ENSG00000109846	CRYAB	146.40606	104.3454	0.005
13	ENSG00000070061	ELP1	137.76027	8.727036	0.005236
14	ENSG00000197265	GTF2E2	131.08964	12.08875	0.005
15	ENSG00000172349	IL16	125.01415	2.10021	0.005
16	ENSG00000131408	NR1H2	105.84597	7.703012	0.005
17	ENSG00000171132	PRKCE	103.07024	1.870311	0.005
18	ENSG00000111674	ENO2	101.79179	14.54598	0.005
19	ENSG00000183765	CHEK2	95.229589	12.38366	0.005
20	ENSG00000148180	GSN	83.66463	14.05611	0.005
21	ENSG00000126934	MAP2K2	75.353403	5.777444	0.005236
22	ENSG00000184634	MED12	72.953482	4.791079	0.005
23	ENSG00000181220	ZNF746	67.588168	2.68988	0.005
24	ENSG00000103168	TAF1C	57.452046	2.234641	0.005
25	ENSG00000107438	PDLIM1	55.015514	49.97263	0.02
26	ENSG00000127445	PIN1	54.804242	6.064614	0.005155

Continued on next page

Table B.1: **Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
27	ENSG00000158869	FCER1G	48.953183	17.90782	0.005
28	ENSG00000149480	MTA2	47.737581	1.044046	0.005
29	ENSG00000197386	HTT	46.86594	21.22442	0.005236
30	ENSG00000103496	STX4	46.708316	14.65071	0.015
31	ENSG00000171700	RGS19	45.073173	1.215949	0.005
32	ENSG00000113140	SPARC	43.831426	13.14772	0.005
33	ENSG00000165304	MELK	43.044883	9.925775	0.005236
34	ENSG00000109103	UNC119	42.229918	17.00491	0.041885
35	ENSG00000166908	PIP4K2C	40.160629	22.37643	0.005
36	ENSG00000130713	EXOSC2	28.514143	20.76501	0.005
37	ENSG00000071655	MBD3	28.04492	2.339941	0.005
38	ENSG00000151849	CENPJ	28.011625	6.079306	0.005236
39	ENSG00000167601	AXL	27.116727	8.841844	0.005236
40	ENSG00000150760	DOCK1	26.782633	7.742055	0.005
41	ENSG00000002822	MAD1L1	26.292489	6.150542	0.005236
42	ENSG00000165140	FBP1	25.566658	38.30417	0.005236
43	ENSG00000108854	SMURF2	24.52682	13.81009	0.025
44	ENSG00000002016	RAD52	23.73388	14.02295	0.005
45	ENSG00000149091	DGKZ	22.508349	3.367526	0.005
46	ENSG00000079277	MKNK1	21.785955	16.83497	0.005
47	ENSG00000182621	PLCB1	20.18339	9.031744	0.005
48	ENSG00000160294	MCM3AP	19.992472	8.580309	0.005
49	ENSG00000109339	MAPK10	17.250088	17.85521	0.005
50	ENSG00000101400	SNTA1	16.97895	3.241013	0.005
51	ENSG00000101096	NFATC2	16.875814	2.284409	0.005236
52	ENSG00000135047	CTSL	16.810349	41.42689	0.005
53	ENSG00000116830	TTF2	16.223873	15.27008	0.005236
54	ENSG00000136936	XPA	16.13043	5.322963	0.005
55	ENSG00000146232	NFKBIE	15.328779	8.028467	0.02
56	ENSG00000104490	NCALD	15.009837	19.39583	0.005
57	ENSG00000104365	IKBKB	13.722595	16.2571	0.005236
58	ENSG00000114739	ACVR2B	13.430946	2.026362	0.005
59	ENSG00000196313	POM121	12.325685	4.433905	0.005236
60	ENSG00000105647	PIK3R2	11.368366	3.261913	0.005
61	ENSG00000132664	POLR3F	11.205922	8.211849	0.01

Continued on next page

Table B.1: **Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
62	ENSG00000180530	NRIP1	11.199377	11.37618	0.005236
63	ENSG00000205250	E2F4	11.02081	7.398753	0.005
64	ENSG00000101017	CD40	10.292378	4.942609	0.005
65	ENSG00000196154	S100A4	10.169565	48.03548	0.005
66	ENSG00000186318	BACE1	9.657818	4.380395	0.005
67	ENSG00000066117	SMARCD1	9.3683661	3.478998	0.005236
68	ENSG00000134982	APC	9.2922105	13.05558	0.005
69	ENSG00000172354	GNB2	9.2880539	7.935466	0.005236
70	ENSG00000068028	RASSF1	9.2595008	9.923702	0.005
71	ENSG00000133895	MEN1	8.7508287	1.59018	0.005236
72	ENSG00000109472	CPE	8.5848494	39.94366	0.005
73	ENSG00000071539	TRIP13	8.5208795	3.96534	0.005236
74	ENSG00000082074	FYB1	8.4927284	2.457428	0.005155
75	ENSG00000151422	FER	8.3451181	6.523093	0.005
76	ENSG00000197971	MBP	8.2837445	4.458641	0.005
77	ENSG00000185043	CIB1	7.7948076	150.8311	0.015
78	ENSG00000178104	PDE4DIP	7.7314185	20.31789	0.005
79	ENSG00000135698	MPHOSPH6	7.726587	60.67041	0.005236
80	ENSG00000272325	NUDT3	7.4441298	15.74288	0.005
81	ENSG00000055163	CYFIP2	7.4422544	10.84894	0.005
82	ENSG00000196924	FLNA	7.3470104	6.894572	0.005236
83	ENSG00000070770	CSNK2A2	7.201422	31.81628	0.005
84	ENSG00000063244	U2AF2	7.1385251	2.994136	0.005155
85	ENSG00000184557	SOCS3	6.9993719	2.117139	0.005
86	ENSG00000004142	POLDIP2	6.877592	23.11806	0.005236
87	ENSG00000123143	PKN1	6.8031887	2.901177	0.005
88	ENSG00000185651	UBE2L3	6.8006707	110.0882	0.035
89	ENSG00000107281	NPDC1	6.7156605	7.54119	0.005
90	ENSG00000157933	SKI	6.7080208	0.898388	0.005
91	ENSG00000104814	MAP4K1	6.6054091	7.019274	0.005236
92	ENSG00000076003	MCM6	6.4645929	9.726372	0.005464
93	ENSG00000138443	ABI2	6.3617729	11.09852	0.005
94	ENSG00000112062	MAPK14	6.2969675	15.10033	0.005236
95	ENSG00000143333	RGS16	6.1774093	23.61302	0.005
96	ENSG00000141644	MBD1	5.9671471	6.786761	0.005

Continued on next page

Table B.1: **Top 100 dynamic cancer drivers from HER2time(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
97	ENSG00000069399	BCL3	5.8257944	1.633143	0.041885
98	ENSG00000204463	BAG6	5.8257488	34.55687	0.005
99	ENSG00000173020	GRK2	5.7530544	1.199343	0.005
100	ENSG00000140443	IGF1R	5.6630238	29.15377	0.005

B.4 Top 100 dynamic cancer drivers from VIMtime(SC)

Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact (absolute value - descending order). A full comprehensive list with the 545 drivers inferred from VIMtime(SC) can be found online in *"supplementary table 2 - dynamic cancer drivers VIMtime(SC).csv"* (folder: *inst/Supplementary/*) at `github.com/AndresMCB/DynamicCancerDriver`. Alternatively, it can be accessed at `https://doi.org/10.6084/m9.figshare.20448384`.

Table B.2: **Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact (absolute value - descending order).**

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
1	ENSG00000167601	AXL	967.0142	9.19085	0.005
2	ENSG00000026025	VIM	463.5813	48.21586	0.005
3	ENSG00000171843	MLLT3	176.5936	34.9199	0.005
4	ENSG00000137265	IRF4	147.9162	6.385924	0.005
5	ENSG00000118971	CCND2	126.1112	5.48083	0.005464
6	ENSG00000113739	STC2	113.5235	10.66576	0.005076
7	ENSG00000057657	PRDM1	92.68285	7.910361	0.032787
8	ENSG00000071655	MBD3	82.45974	2.402318	0.005
9	ENSG00000028137	TNFRSF1B	63.52454	2.041571	0.005
10	ENSG00000168078	PBK	63.28292	9.931831	0.005
11	ENSG00000182568	SATB1	60.5641	10.27128	0.005
12	ENSG00000010610	CD4	56.69995	5.148583	0.005464
13	ENSG00000111674	ENO2	55.86033	14.47186	0.005
14	ENSG00000105926	PALS2	55.02897	9.489363	0.005
15	ENSG00000117676	RPS6KA1	54.11415	9.331178	0.005464

Continued on next page

Table B.2: **Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
16	ENSG00000132906	CASP9	50.24631	2.461116	0.005
17	ENSG00000186318	BACE1	50.0033	4.759209	0.005
18	ENSG00000131773	KHDRBS3	38.77092	27.90944	0.016393
19	ENSG00000141456	PELP1	37.1198	2.241423	0.005
20	ENSG00000077684	JADE1	36.06523	4.29893	0.005
21	ENSG00000173757	STAT5B	33.83659	6.183731	0.005
22	ENSG00000111145	ELK3	30.53059	3.918399	0.005464
23	ENSG00000105647	PIK3R2	30.15169	3.444846	0.005
24	ENSG00000137486	ARRB1	29.05979	4.284124	0.005464
25	ENSG00000175104	TRAF6	28.96674	9.011639	0.005464
26	ENSG00000163703	CRELD1	24.05496	10.10235	0.005
27	ENSG00000115414	FN1	23.37931	52.62427	0.005464
28	ENSG00000175220	ARHGAP1	19.80497	78.5678	0.005
29	ENSG00000160972	PPP1R16A	18.59525	4.478329	0.005
30	ENSG00000162236	STX5	18.02857	6.38711	0.005
31	ENSG00000119689	DLST	16.45453	7.33199	0.005464
32	ENSG00000100320	RBFOX2	16.0485	24.79369	0.005
33	ENSG00000003400	CASP10	15.96417	0.997455	0.005464
34	ENSG00000185950	IRS2	13.8896	2.315517	0.005236
35	ENSG00000136936	XPA	13.17041	5.265493	0.005236
36	ENSG00000173020	GRK2	12.7959	1.312553	0.005
37	ENSG00000120696	KBTBD7	12.36385	4.717568	0.005
38	ENSG00000127511	SIN3B	11.59642	2.969183	0.016393
39	ENSG00000146648	EGFR	11.59405	5.714505	0.005
40	ENSG00000127946	HIP1	10.03739	7.027747	0.005
41	ENSG00000055163	CYFIP2	9.944083	11.21787	0.005
42	ENSG00000109103	UNC119	9.843793	15.89718	0.005
43	ENSG00000142166	IFNAR1	9.777363	33.20074	0.005
44	ENSG00000103319	EEF2K	9.547039	7.269116	0.005025
45	ENSG00000211899	IGHM	9.371626	5.401124	0.005
46	ENSG00000105221	AKT2	9.1333	2.720025	0.005
47	ENSG00000145604	SKP2	8.58281	8.252342	0.005
48	ENSG00000105851	PIK3CG	8.454076	0.414084	0.005
49	ENSG00000020633	RUNX3	8.276306	4.934203	0.005
50	ENSG00000107281	NPDC1	7.478224	7.652843	0.005236

Continued on next page

Table B.2: **Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
51	ENSG00000131323	TRAF3	7.043142	2.680371	0.015
52	ENSG00000126351	THRA	6.92421	6.414458	0.005
53	ENSG00000149782	PLCB3	6.329103	0.876288	0.005236
54	ENSG00000091073	DTX2	6.300156	6.078588	0.005
55	ENSG00000101189	MRGBP	6.244981	5.324855	0.032787
56	ENSG00000118785	SPP1	6.167655	53.54017	0.005464
57	ENSG00000114739	ACVR2B	6.003658	1.868845	0.005236
58	ENSG00000069974	RAB27A	5.893092	4.606596	0.005464
59	ENSG00000002822	MAD1L1	5.814716	5.441066	0.005
60	ENSG00000154710	RABGEF1	5.579696	13.64613	0.005
61	ENSG00000087077	TRIP6	5.547595	5.577101	0.005236
62	ENSG00000129675	ARHGEF6	5.439333	5.522459	0.005464
63	ENSG00000097007	ABL1	5.432701	3.58441	0.005155
64	ENSG00000163513	TGFBR2	5.271851	2.315112	0.005464
65	ENSG00000157873	TNFRSF14	5.147366	3.952381	0.005076
66	ENSG00000108599	AKAP10	5.012705	12.19545	0.005
67	ENSG00000090061	CCNK	4.976074	16.32538	0.005
68	ENSG00000134909	ARHGAP32	4.867756	11.70782	0.005
69	ENSG00000182979	MTA1	4.842439	0.738495	0.005
70	ENSG00000168884	TNIP2	4.80785	2.693697	0.005
71	ENSG00000113140	SPARC	4.789958	11.11228	0.005
72	ENSG00000149311	ATM	4.789085	11.01685	0.005236
73	ENSG00000110713	NUP98	4.773384	38.32559	0.005
74	ENSG00000090104	RGS1	4.718807	71.36185	0.005
75	ENSG00000151276	MAGI1	4.455276	4.736045	0.005464
76	ENSG00000104365	IKBKB	4.432541	14.23666	0.005
77	ENSG00000165029	ABCA1	4.128617	4.094829	0.005
78	ENSG00000121931	LRIF1	4.083649	16.02418	0.005236
79	ENSG00000165140	FBP1	3.915504	31.62359	0.005
80	ENSG00000125952	MAX	3.834127	10.63444	0.038251
81	ENSG00000176903	PNMA1	3.726582	3.425579	0.005
82	ENSG00000143622	RIT1	3.681624	50.47222	0.005
83	ENSG00000096968	JAK2	3.587398	1.709275	0.015707
84	ENSG00000132485	ZRANB2	3.561511	30.98357	0.005
85	ENSG00000077782	FGFR1	3.528573	30.96438	0.005

Continued on next page

Table B.2: **Top 100 dynamic cancer drivers from VIMtime(SC), ranked by Relative Causal Impact (absolute value - descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p.val
86	ENSG00000160310	PRMT2	3.285659	29.13423	0.005236
87	ENSG00000153071	DAB2	3.205532	4.340487	0.005
88	ENSG00000077942	FBLN1	3.194051	9.569679	0.005
89	ENSG00000170542	SERPINB9	3.154632	29.79213	0.005
90	ENSG00000135966	TGFBRAP1	3.123153	3.721607	0.005
91	ENSG00000157110	RBPMS	3.07377	14.16062	0.005
92	ENSG00000082781	ITGB5	2.980313	18.00619	0.005
93	ENSG00000114316	USP4	2.943463	24.49412	0.005
94	ENSG00000103168	TAF1C	2.938426	1.694229	0.005
95	ENSG00000095794	CREM	2.900339	38.05375	0.005236
96	ENSG00000117318	ID3	2.728742	38.57592	0.005
97	ENSG00000091831	ESR1	2.728374	37.22242	0.02
98	ENSG00000178209	PLEC	2.710291	1.874952	0.005
99	ENSG00000169016	E2F6	2.699985	8.172608	0.045
100	ENSG00000127824	TUBA4A	2.687856	31.44567	0.005

B.5 Stratification of inferred driver genes (intersection set) based on literature support

Stratified list of the 215 genes discovered as drivers in the intersection set of drivers inferred from both pseudotime orders (VIMtime(SC) and Her2time(SC)). Stratification was done by analysing 333 references obtained with the following the inclusion criteria: (i) paper explicitly refers the gene has a significant role in cancer, (ii) publication dated 2016 onwards and (iii) study is published in a journal with impact factor 5.1 or higher (as shown in web of knowledge, website visited in August 2021).

Inferred DCDs are categorised as follows: **A** - the gene is a CGC gene, **B** - the gene has literature support as influential for breast cancer, **C** - the gene have literature support as influential for other cancer but breast cancer, and **D** - there is not literature describing the gene as influential for cancer.

Columns “Experimental support breast cancer” and “Experimental support other

cancer" refers to the references analysed for stratification. List of references used for stratification can be found online in *"supplementary table 3 - DCD stratification.xlsx"* (folder: *inst/Supplementary/*) at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.**

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
1	AXL	967.0142014	0.005	1,3	2,4	B
2	IRF4	147.9161761	0.005			A
3	STC2	113.5235274	0.005076142		5,6	C
4	MBD3	82.45974119	0.005		7	C
5	IGF1R	76.20768454	0.005	8,10,11	9,12,13,14,15	B
6	TNFRSF1B	63.52454174	0.005	NA	NA	D
7	SATB1	60.56410376	0.005		16,17	C
8	CD4	56.69994808	0.005464481	NA	NA	D
9	ENO2	55.860335	0.005		18	C
10	RPS6KA1	54.11415281	0.005464481	NA	NA	D
11	CASP9	50.24631379	0.005			A
12	BACE1	50.00329742	0.005	NA	NA	D
13	JADE1	36.06522921	0.005	NA	NA	D
14	STAT5B	33.83658622	0.005			A
15	CRYAB	33.75630913	0.005		19, 20	C
16	PIK3R2	30.15169315	0.005	30	29	B
17	TRAF6	28.96673637	0.005464481	25,28	21, 22, 23, 24, 26, 27	B
18	A2M	27.6610551	0.005	NA	31	C
19	CRELD1	24.05495593	0.005	NA	NA	D
20	XPA	13.17040735	0.005235602			A
21	GRK2	12.79589505	0.005	34,35	32,33	B
22	EGFR	11.59404874	0.005			A
23	CYFIP2	9.944082861	0.005	NA	NA	D
24	UNC119	9.843793394	0.005		36	C
25	RUNX3	8.276305563	0.005	42	37, 38, 39, 40, 41, 43	B
26	NPDC1	7.478223516	0.005235602	NA	NA	D

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
27	THRA	6.924210436	0.005	NA	NA	D
28	DTX2	6.300155733	0.005	NA	NA	D
29	SPP1	6.167654995	0.005464481	50	44,45, 46, 47, 48, 49	B
30	ACVR2B	6.003658115	0.005235602	NA	NA	D
31	MAD1L1	5.814715614	0.005		51	C
32	ASH2L	5.645810475	0.027322404		52,53,54	C
33	RABGEF1	5.579695989	0.005	NA	55	C
34	ABL1	5.432701247	0.005154639			A
35	TGFBR2	5.271851301	0.005464481			A
36	TNFRSF14	5.147365967	0.005076142			A
37	AKAP10	5.012704774	0.005	NA	NA	D
38	SPARC	4.789957839	0.005	60, 62, 63	56, 57, 58, 59, 61	B
39	NUP98	4.773384443	0.005			A
40	RGS1	4.718806771	0.005	65	64	B
41	IKBKB	4.432541012	0.005			A
42	LRIF1	4.083649021	0.005235602	NA	NA	D
43	FBP1	3.915504317	0.005	68, 70	66, 67, 69, 71	B
44	RIT1	3.681624363	0.005		72	C
45	FGFR1	3.528573393	0.005			A
46	MAPT	3.365975927	0.032786885	73	74, 75	B
47	SERPINB9	3.154632126	0.005		76, 77, 78	C
48	RBPMS	3.073769752	0.005		79, 80	C
49	TAF1C	2.938426219	0.005	NA	NA	D
50	ESR1	2.728373677	0.02			A
51	E2F6	2.699985221	0.045		81, 82, 83, 84, 85	C
52	SNUPN	2.673898011	0.005		86	C
53	ZNF76	2.538945656	0.005	NA	NA	D
54	MCF2L	2.472573458	0.005		87	C
55	DLG5	2.24600923	0.005	88	89	B
56	BNIP3L	2.142233904	0.005235602		90, 91, 92, 93	C

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
57	FBXW7	1.947894421	0.005			A
58	PTPRA	1.861177757	0.005		94, 95	C
59	TAF10	1.837143758	0.005235602	NA	NA	D
60	MAPK10	1.796315656	0.04		96, 97	C
61	S100A4	1.782618039	0.005	99	98, 100, 101, 102	B
62	TAF6	1.739377518	0.005	NA	NA	D
63	NRAS	1.680597763	0.005			A
64	CDK14	1.661580434	0.005235602	108	103, 104, 105, 106, 107	B
65	PLK1	1.647920019	0.005235602	109, 111	110, 112, 113	B
66	AKAP1	1.635375081	0.005		114	C
67	MDC1	1.549019601	0.005	115, 116, 118	117	B
68	FUNDC2	1.523157972	0.04	NA	NA	D
69	EID1	1.403248691	0.005154639	NA	NA	D
70	ABCF3	1.387964445	0.005	NA	NA	D
71	PRSS23	1.33520973	0.005		119	C
72	TPD52L1	1.318746991	0.005		120	C
73	IGFBP5	1.308922954	0.015	122, 123	121, 124	B
74	SGK1	1.301694331	0.01			A
75	KHSRP	1.282361211	0.02		125, 126, 127	C
76	PSMF1	1.251299412	0.01		128	C
77	CHMP3	1.234367554	0.005	129		B
78	BAG1	1.206010201	0.005		130, 131	C
79	TSC1	1.180396291	0.038251366			A
80	ELP1	1.115676291	0.015		132, 133, 134	C
81	BLOC1S2	1.038291694	0.015873016	NA	NA	D
82	CTSB	0.988372989	0.005	135	136	B
83	EXOSC10	0.964985943	0.005235602		137	C
84	OGT	0.962672835	0.005235602	142	138, 139, 140, 141	B
85	EIF3E	0.940256033	0.005			A
86	CD44	0.895329581	0.010471204	143	144, 145, 146	B

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
87	PARD6B	0.870545389	0.005	NA	NA	D
88	HTRA2	0.848661098	0.005235602		147, 148	C
89	ARHGEF11	0.846322758	0.005	149		B
90	EXOSC2	0.843285251	0.02		150	C
91	APP	0.825156282	0.005	151	152	B
92	HNRNPUL1	0.81804556	0.005		153, 154	C
93	GRB7	0.809019429	0.005	159	155, 156, 157, 158	B
94	POLR3F	0.793511342	0.005	NA	NA	D
95	CDK5RAP2	0.790118219	0.005	NA	NA	D
96	HOXB7	0.789145389	0.005	162	160, 161, 163, 164	B
97	LUC7L2	0.776363024	0.035	NA	NA	D
98	ECH1	0.771615027	0.031413613		165	C
99	UBQLN4	0.766980639	0.005		166, 167	C
100	SUFU	0.739693903	0.025			A
101	ATF1	0.738473161	0.005			A
102	DAZAP2	0.731282775	0.025		168	C
103	COL1A1	0.73037602	0.037634409			A
104	MDM2	0.726194087	0.005235602			A
105	SAFB	0.725460456	0.005		169	C
106	HSF2	0.710555083	0.025		170, 171	C
107	ITSN1	0.710178024	0.005		172	C
108	TIAM1	0.705877341	0.010471204	173, 174	175, 176, 177	B
109	CDKN2A	0.70480004	0.005			A
110	TAF4	0.689749116	0.02	NA	NA	D
111	HDAC3	0.688670849	0.01	181, 182	178, 179, 180	B
112	NFYA	0.684840661	0.015706806	186	183, 184, 185	B
113	MAFG	0.676020057	0.005235602	191	187, 188, 189, 190	B
114	SNW1	0.667004501	0.02	NA	NA	D
115	RAB4A	0.662468845	0.02617801	193, 194	192	B
116	IRAK1	0.660124242	0.02		195, 196, 197	C

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
117	CDK9	0.658898753	0.025	201	198, 199, 200	B
118	KRT19	0.65482944	0.030927835	203, 204	202, 205	B
119	RHOQ	0.645667342	0.01	NA	NA	D
120	NFKBIA	0.645665593	0.005		206	C
121	SAP30	0.645505384	0.015706806	NA	NA	D
122	ZNF250	0.645488584	0.031413613	NA	NA	D
123	COPS2	0.645385983	0.005235602		207, 208	C
124	SOS1	0.643451256	0.015	212	209, 210, 211	B
125	NRIP1	0.642793293	0.04		213, 214	C
126	MED8	0.642262712	0.010471204		215	C
127	TFE3	0.633390383	0.025			A
128	NEDD9	0.629723686	0.02	218	216, 217	B
129	MCM5	0.626311904	0.02		219,220	C
130	VASP	0.61727742	0.020942408	222, 223, 224	221	B
131	SERTAD1	0.616910583	0.005		225	C
132	USP7	0.607796294	0.005	229, 230	226, 227, 228	B
133	PTPRS	0.607229286	0.005		231	C
134	C1QBP	0.607110493	0.005		232	C
135	GAK	0.606798655	0.02		233	C
136	CBX4	0.605939368	0.01	236	234, 235	B
137	POU2F1	0.594122977	0.047120419		237, 238, 239, 240	C
138	CCT7	0.582703372	0.005		241	C
139	HCFC1	0.582185614	0.010471204		242, 243	C
140	HNRNPD	0.579503775	0.005		244	C
141	TUBB	0.575174893	0.005		245	C
142	MAP2K2	0.573638167	0.03			A
143	EIF2AK2	0.568282917	0.005	248	246, 247	B
144	HTATSF1	0.567040582	0.005		249	C
145	STX4	0.552911154	0.015706806	NA	NA	D
146	KDM5A	0.550021348	0.005			A
147	RAB2A	0.545077557	0.005	251	250	B
148	CYTH2	0.53904391	0.020942408	NA	NA	D

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
149	CHMP1B	0.538855309	0.005235602	NA	NA	D
150	GTF2B	0.536917285	0.031413613	NA	NA	D
151	CRK	0.521047235	0.005	254, 255, 256, 257	252, 253, 253	B
152	ARHGAP5	0.511172645	0.025			A
153	PDIA3	0.505583644	0.02		258	C
154	HRAS	0.50422088	0.026315789			A
155	LNX1	0.503131054	0.01		259, 260	C
156	SIN3A	0.502289444	0.016393443	261, 262, 263		B
157	REPS1	0.501355739	0.005	NA	NA	D
158	MBIP	0.493013952	0.02617801		264	C
159	SORBS2	0.485365014	0.005154639	268	265, 266, 267	B
160	SNX6	0.479337084	0.005	NA	NA	D
161	DDIT3	0.47874927	0.005			A
162	MCM3AP	0.478400231	0.031413613	NA	NA	D
163	NAGK	0.473046255	0.015	NA	NA	D
164	ZMYND8	0.459500373	0.015	270, 271	269, 272	B
165	ZYX	0.456377826	0.005235602		273, 274	C
166	POM121	0.450842532	0.01		275	C
167	VDR	0.448824608	0.03	277, 278	276	B
168	SF3B3	0.445194067	0.015	280	279, 281	B
169	CFL1	0.441752005	0.005	284	282, 283	B
170	CEBPZ	0.434346524	0.01	NA	NA	D
171	TGS1	0.431577168	0.015	NA	NA	D
172	HES1	0.43137205	0.005	286	285, 287	B
173	CDH1	0.427097861	0.015			A
174	NCL	0.42612716	0.005	289	288, 290	B
175	ELF1	0.423943634	0.010471204		291,292	C
176	SNAPIN	0.418197689	0.02	NA	NA	D
177	SFPQ	0.409519091	0.005235602			A
178	MBD2	0.408656038	0.015	293	294, 295	B
179	NRBP1	0.395058879	0.02		296	C
180	ELOC	0.391606689	0.04	NA	NA	D

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
181	FUBP1	0.379739643	0.01			A
182	PPHLN1	0.378460004	0.005	NA	NA	D
183	ERH	0.376826494	0.02		297, 298	C
184	EIF4G2	0.363380473	0.005235602		299, 300, 301	C
185	ATP5F1B	0.363340419	0.005	NA	NA	D
186	GOLM1	0.357810743	0.03	304	302, 303, 305	B
187	ZMYND11	0.35700878	0.03		306	C
188	XPO1	0.351155062	0.035			A
189	ZNF165	0.350704685	0.036649215	307		B
190	RBBP7	0.349586315	0.02	NA	NA	D
191	TADA3	0.347982976	0.035	NA	NA	D
192	SYNCRIP	0.342508307	0.005	309	308	B
193	CTNNA1	0.340148812	0.02	312	310, 311	B
194	RERE	0.333554408	0.005	NA	NA	D
195	EXOSC8	0.327604646	0.005		313	C
196	TAPBP	0.326985751	0.02	NA	NA	D
197	PTEN	0.326903206	0.045			A
198	ACTR2	0.323148191	0.02		314	C
199	PLEKHF2	0.32017891	0.045	NA	NA	D
200	YWHAZ	0.316173669	0.005235602	315	316, 317	B
201	PIK3R3	0.311972103	0.025		318	C
202	ATP1B1	0.307474317	0.02		319	C
203	CBX3	0.299992512	0.031413613	236	320, 321	B
204	B2M	0.299314308	0.03			A
205	PEBP1	0.277198238	0.01	322		B
206	LSM5	0.265822276	0.005235602	NA	NA	D
207	COPS5	0.253249674	0.025	323, 324	325	B
208	NAP1L1	0.246656963	0.015706806		326, 327, 328	C
209	DBI	0.240972389	0.025	NA	NA	D
210	SNRPD2	0.239353585	0.045	NA	NA	D
211	RPS27A	0.229363231	0.005	NA	NA	D
212	STAT3	0.2235875	0.02			A
213	NPM1	0.209031677	0.015			A

Continued on next page

Table B.3: **Stratification of inferred driver genes (intersection set) based on literature support.** (Continued)

	HGNC.symbol	RelEffect (abs. value)	p.val	Experimental support breast cancer	Experimental support other cancer	Category
214	SERBP1	0.193195775	0.04		329	C
215	GAPDH	0.186816582	0.03		330, 331, 332, 333	C

B.6 Top 10 GO terms HER2time(SC)

Analysis performed by using enrichr (https://maayanlab.cloud/Enrichr/) The full outcome of the analysis (including the list of enriched genes) can be found online in *"Supplementary/supplementary table 4 - GO terms HER2time(SC).txt"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.4: **Top 10 GO terms HER2time(SC) ranked by p-value.**

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
positive regulation of transcription, DNA-templated (GO:0045893)	144/1183	2.06E-49	7.43E-46	5.530837	620.0322
positive regulation of transcription by RNA polymerase II (GO:0045944)	114/908	6.47E-40	1.17E-36	5.450645	491.8462
regulation of transcription by RNA polymerase II (GO:0006357)	173/2206	1.01E-33	1.21E-30	3.428122	260.4697
regulation of transcription, DNA-templated (GO:0006355)	172/2244	2.90E-32	2.62E-29	3.328918	241.7361
negative regulation of transcription, DNA-templated (GO:0045892)	102/948	7.19E-30	5.18E-27	4.455229	298.9702
regulation of apoptotic process (GO:0042981)	86/742	2.02E-27	1.22E-24	4.742796	291.5191

Continued on next page

Table B.4: **Top 10 GO terms HER2time(SC) ranked by p-value.** (Continued)

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
positive regulation of nucleic acid-templated transcription (GO:1903508)	70/511	1.06E-26	5.47E-24	5.634326	336.9725
negative regulation of programmed cell death (GO:0043069)	59/381	2.11E-25	9.50E-23	6.412707	364.366
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	70/547	6.56E-25	2.63E-22	5.199201	289.5087
negative regulation of apoptotic process (GO:0043066)	65/485	2.70E-24	9.74E-22	5.448538	295.6848

B.7 Top 10 GO terms VIMtime(SC)

Analysis performed by using enrichr (https://maayanlab.cloud/Enrichr/) The full outcome of the analysis (including the list of enriched genes) can be found online in *"Supplementary/supplementary table 5 - GO terms VIMtime(SC).txt"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.5: **Top 10 GO terms VIMtime(SC) ranked by p-value.**

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
positive regulation of transcription, DNA-templated (GO:0045893)	125/1183	6.05E-41	1.98E-37	5.17514	479.2502
regulation of transcription by RNA polymerase II (GO:0006357)	154/2206	2.28E-29	3.73E-26	3.34034	220.3038
positive regulation of transcription by RNA polymerase II (GO:0045944)	93/908	4.79E-29	4.97E-26	4.705793	306.8578
negative regulation of transcription, DNA-templated (GO:0045892)	95/948	6.06E-29	4.97E-26	4.603856	299.1244
regulation of transcription, DNA-templated (GO:0006355)	152/2244	1.60E-27	1.05E-24	3.210067	198.061
regulation of apoptotic process (GO:0042981)	78/742	4.64E-25	2.54E-22	4.726716	264.84
negative regulation of transcription by RNA polymerase II (GO:0000122)	73/684	7.26E-24	3.40E-21	4.769938	254.1402
negative regulation of cellular macromolecule biosynthetic process (GO:2000113)	61/547	5.35E-21	2.19E-18	4.919179	229.6169
regulation of intracellular signal transduction (GO:1902531)	51/437	1.53E-18	5.57E-16	5.10016	209.2216

Continued on next page

Table B.5: **Top 10 GO terms VIMtime(SC) ranked by p-value.** (Continued)

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
negative regulation of nucleic acid-templated transcription (GO:1903507)	51/464	1.99E-17	6.53E-15	4.759987	183.047

B.8 Top 10 Gene-Disease Associated terms (GDAs) - HER2Time(SC)

Analysis performed by using enrichr (https://maayanlab.cloud/Enrichr/) The full outcome of the analysis (including the list of enriched genes) can be found online in *"Supplementary/supplementary table 6 - GDAs HER2time(SC).txt"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.6: **Top 10 GDAs HER2Time(SC) ranked by p-value.**

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
Mammary Neoplasms	244/2387	1.71E-73	1.12E-69	5.456696	914.2905
Malignant neoplasm of breast	353/5054	4.62E-69	1.52E-65	4.396229	691.7339
Breast Carcinoma	348/4963	4.75E-68	1.04E-64	4.353829	674.9187
Neoplasm Metastasis	292/3920	8.59E-59	1.41E-55	4.067594	543.8443
Malignant neoplasm of prostate	262/3239	3.30E-58	4.34E-55	4.225159	559.2211
Tumor Progression	203/2090	2.45E-55	2.68E-52	4.697222	590.6636
Carcinogenesis	291/4065	1.18E-54	1.11E-51	3.848428	477.8681
Prostate carcinoma	249/3145	9.09E-53	7.47E-50	3.996284	478.8745
Malignant neoplasm of pancreas	180/1846	1.85E-48	1.35E-45	4.517939	496.5704
Non-Small Cell Lung Carcinoma	199/2243	8.76E-48	5.76E-45	4.171255	451.9703

B.9 Top 10 Gene-Disease Associated terms (GDAs) - VIMTime(SC)

Analysis performed by using enrichr (https://maayanlab.cloud/Enrichr/) The full outcome of the analysis (including the list of enriched genes) can be found in *"Supplementary/ supplementary table 7 - GDAs VIMtime(SC).txt"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.7: **Top 10 GDAs VIMTime(SC) ranked by p-value.**

Term	Overlap	p.val	Adjusted p.val	Odds Ratio	Combined Score
Carcinogenesis	273/4065	1.39E-55	8.61E-52	4.145724	523.6669
Malignant neoplasm of breast	306/5054	1.64E-54	5.08E-51	3.965856	491.1573
Breast Carcinoma	301/4963	2.31E-53	4.79E-50	3.914359	474.4149
Malignant neoplasm of prostate	236/3239	3.05E-52	4.74E-49	4.184243	496.3296
Neoplasm Metastasis	262/3920	3.88E-52	4.81E-49	3.998027	473.2837
Prostate carcinoma	228/3145	1.27E-49	1.32E-46	4.077765	459.0921
Malignant neoplasm of ovary	174/2026	8.65E-46	7.67E-43	4.457805	462.5485
Mammary Neoplasms	190/2387	1.11E-45	8.59E-43	4.204222	435.1976
Liver carcinoma	236/3593	4.25E-44	2.93E-41	3.66247	365.7586
Malignant neoplasm of pancreas	162/1846	1.94E-43	1.21E-40	4.463608	438.9855

B.10 Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA.

Tie break by Relative Causal Impact (descending order). A full comprehensive list with the drivers inferred from HER2time(Bulk) can be found in *"Supplementary/ supplementary table 8 - dynamic cancer drivers HER2time(Bulk).csv"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.8: **Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA (descending order).**

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
1	ENSG00000121879	PIK3CA	328	0.245815	570.8735	0.005
2	ENSG00000151914	DST	75	-0.52536	-18758.5	0.005155
3	ENSG00000198947	DMD	59	-0.67595	-2818.27	0.015
4	ENSG00000039068	CDH1	55	1.973177	20281.88	0.005
5	ENSG00000141027	NCOR1	44	0.298908	1743.139	0.005
6	ENSG00000127603	MACF1	43	0.961934	10218.31	0.005236
7	ENSG00000084674	APOB	42	3.879672	16.24728	0.005
8	ENSG00000081479	LRP2	40	5.566193	6432.024	0.005
9	ENSG00000085224	ATRX	39	0.520398	2762.932	0.005
10	ENSG00000196712	NF1	36	3.157311	6445.952	0.005
11	ENSG00000166147	FBN1	34	0.54127	6134.835	0.005181
12	ENSG00000171862	PTEN	34	0.451308	2642.034	0.005
13	ENSG00000165124	SVEP1	33	-0.81069	-6400.35	0.025
14	ENSG00000111642	CHD4	32	-0.09729	-2167.03	0.025
15	ENSG00000152818	UTRN	30	0.341521	1549.364	0.015464
16	ENSG00000198734	F5	30	-0.53017	-306.427	0.015
17	ENSG00000141736	ERBB2	28	19.25144	53328.42	0.005
18	ENSG00000065559	MAP2K4	28	0.434703	712.4424	0.005
19	ENSG00000196924	FLNA	27	0.434926	25156.14	0.005236
20	ENSG00000142208	AKT1	27	-0.66189	-29272.6	0.015
21	ENSG00000104517	UBR5	26	0.289933	2360.438	0.015
22	ENSG00000149311	ATM	25	0.319154	705.318	0.005236
23	ENSG00000178209	PLEC	24	0.635678	10220.11	0.005236
24	ENSG00000101680	LAMA1	24	-0.83253	-1963.97	0.01
25	ENSG00000178104	PDE4DIP	23	0.695621	2178.394	0.010471
26	ENSG00000118058	KMT2A	23	0.438644	1542.594	0.005
27	ENSG00000065534	MYLK	23	-0.64402	-13445.6	0.005181
28	ENSG00000130396	AFDN	22	1.759308	4982.454	0.015
29	ENSG00000278540	ACACA	22	1.11172	3466.663	0.005076
30	ENSG00000142798	HSPG2	22	0.527988	6548.406	0.005236
31	ENSG00000137076	TLN1	22	0.420664	7233.55	0.015707
32	ENSG00000167522	ANKRD11	22	0.120779	506.163	0.025
33	ENSG00000196367	TRRAP	22	-0.32451	-3181.78	0.005076
34	ENSG00000203710	CR1	22	-0.40821	-65.5736	0.035

Continued on next page

Table B.8: **Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA (descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
35	ENSG00000079102	RUNX1T1	22	-0.5674	-488.375	0.025381
36	ENSG00000007314	SCN4A	21	2.716321	93.06517	0.005
37	ENSG00000169946	ZFPM2	21	2.087188	277.956	0.005236
38	ENSG00000159216	RUNX1	21	0.547964	4085.788	0.005
39	ENSG00000134250	NOTCH2	21	0.289734	3291.523	0.026178
40	ENSG00000079841	RIMS1	20	18.69365	186.2013	0.005
41	ENSG00000029534	ANK1	20	3.152989	354.0884	0.005236
42	ENSG00000001626	CFTR	20	2.695465	45.47265	0.005
43	ENSG00000138162	TACC2	20	2.164844	2471.161	0.005
44	ENSG00000110799	VWF	20	1.022028	5815.601	0.005076
45	ENSG00000142949	PTPRF	20	0.964981	22105.08	0.005
46	ENSG00000197386	HTT	20	-0.35377	-3898.92	0.015464
47	ENSG00000225830	ERCC6	20	-0.52358	-1004.93	0.005
48	ENSG00000197565	COL4A6	19	679.3059	660.4388	0.005
49	ENSG00000107186	MPDZ	19	1.05736	1087.061	0.005
50	ENSG00000273079	GRIN2B	19	-0.83304	-115.878	0.035533
51	ENSG00000150995	ITPR1	19	-4.6251	6397.099	0.005348
52	ENSG00000188153	COL4A5	19	-122.31	3590.755	0.005
53	ENSG00000081052	COL4A4	18	0.683564	75.82118	0.005
54	ENSG00000012048	BRCA1	18	0.469256	395.4063	0.015
55	ENSG00000064012	CASP8	18	0.451831	502.2222	0.036082
56	ENSG00000131626	PPFIA1	18	0.434158	1626.502	0.005076
57	ENSG00000107485	GATA3	18	-27.3602	38154.09	0.005
58	ENSG00000174469	CNTNAP2	17	65.65806	1993.775	0.005
59	ENSG00000147255	IGSF1	17	29.03406	1806.542	0.005236
60	ENSG00000185008	ROBO2	17	3.575056	554.7719	0.015707
61	ENSG00000146648	EGFR	17	1.729244	3156.434	0.005
62	ENSG00000018236	CNTN1	17	1.566935	485.53	0.015
63	ENSG00000112282	MED23	17	1.517827	2351.494	0.005
64	ENSG00000170776	AKAP13	17	0.900338	4075.433	0.005
65	ENSG00000197081	IGF2R	17	0.300825	2326.957	0.005236
66	ENSG00000100393	EP300	17	0.2657	1560.264	0.045
67	ENSG00000102974	CTCF	17	0.189471	715.382	0.01
68	ENSG00000132694	ARHGEF11	17	-0.23254	-1560.98	0.015

Continued on next page

Table B.8: **Top 100 dynamic cancer drivers from HER2time(Bulk), ranked by frequency of mutation in TCGA (descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
69	ENSG00000153201	RANBP2	17	-0.35645	-6457.65	0.025
70	ENSG00000178568	ERBB4	16	273.4745	2705.585	0.005
71	ENSG00000089250	NOS1	16	2.853439	14.08916	0.005
72	ENSG00000198753	PLXNB3	16	2.332504	1045.554	0.005236
73	ENSG00000141367	CLTC	16	0.809977	19211.59	0.005
74	ENSG00000167306	MYO5B	16	0.790746	2490.094	0.005
75	ENSG00000197694	SPTAN1	16	0.513481	8191.207	0.005236
76	ENSG00000114127	XRN1	16	0.427979	809.9177	0.005
77	ENSG00000049759	NEDD4L	16	0.361481	1342.941	0.005
78	ENSG00000111859	NEDD9	16	-0.34058	-1786.02	0.025
79	ENSG00000196313	POM121	16	-0.40287	-1549.11	0.04
80	ENSG00000169855	ROBO1	16	-0.45522	-2649.16	0.015
81	ENSG00000077522	ACTN2	16	-0.54663	-159.909	0.04
82	ENSG00000182901	RGS7	15	109.5107	41.92506	0.005236
83	ENSG00000183454	GRIN2A	15	86.18357	125.366	0.015707
84	ENSG00000078328	RBFOX1	15	12.53668	5.686938	0.005208
85	ENSG00000155511	GRIA1	15	6.868589	253.2287	0.005236
86	ENSG00000174175	SELP	15	1.223755	283.5931	0.005
87	ENSG00000112818	MEP1A	15	0.986936	3.173108	0.01
88	ENSG00000132849	PATJ	15	0.95111	3079.854	0.015
89	ENSG00000151276	MAGI1	15	0.408507	675.4461	0.005
90	ENSG00000130985	UBA1	15	-0.23857	-8362.73	0.005
91	ENSG00000196083	IL1RAP	15	-0.48773	-762.174	0.005155
92	ENSG00000197535	MYO5A	15	-0.68703	-5707.46	0.005181
93	ENSG00000133124	IRS4	14	412.3778	117.6291	0.005025
94	ENSG00000148053	NTRK2	14	3.167965	3245.432	0.005236
95	ENSG00000065618	COL17A1	14	2.123479	1853.503	0.005076
96	ENSG00000182752	PAPPA	14	2.048972	161.1572	0.005236
97	ENSG00000106571	GLI3	14	0.827262	1776.633	0.005
98	ENSG00000100852	ARHGAP5	14	0.727277	2731.178	0.005236
99	ENSG00000124151	NCOA3	14	0.684853	1772.362	0.005
100	ENSG00000112851	ERBIN	14	0.527422	3142.484	0.005181

B.11 Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA.

Tie break by Relative Causal Impact (descending order). A full comprehensive list with the drivers inferred from VIMtime(Bulk) can be found in *"supplementary table 9 - dynamic cancer drivers VIMtime(Bulk).csv"* at github.com/AndresMCB/DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.9: **Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA (descending order).**

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
1	ENSG00000121879	PIK3CA	328	0.345788	744.0998	0.005
2	ENSG00000198626	RYR2	85	4.207954	121.5235	0.005
3	ENSG00000151914	DST	75	0.810778	7603.924	0.005
4	ENSG00000163554	SPTA1	57	0.843172	12.33052	0.005025
5	ENSG00000039068	CDH1	55	1.04397	15585.82	0.005
6	ENSG00000127603	MACF1	43	1.600195	12841.21	0.005
7	ENSG00000084674	APOB	42	37.69888	19.92137	0.005
8	ENSG00000081479	LRP2	40	2.038013	5082.172	0.005
9	ENSG00000085224	ATRX	39	0.359272	2133.592	0.005236
10	ENSG00000127914	AKAP9	37	-0.66225	-21957.5	0.005
11	ENSG00000196712	NF1	36	5.059854	7075.278	0.005236
12	ENSG00000065526	SPEN	35	1.822539	4250.23	0.005
13	ENSG00000166147	FBN1	34	7.633048	15482.57	0.005
14	ENSG00000165124	SVEP1	33	1.062377	773.841	0.005
15	ENSG00000111642	CHD4	32	0.205198	3429.26	0.025
16	ENSG00000152818	UTRN	30	9.464905	5506.286	0.005
17	ENSG00000198734	F5	30	2.377713	193.7885	0.02
18	ENSG00000153707	PTPRD	30	-2.2059	1508.761	0.04
19	ENSG00000123384	LRP1	29	9.982512	24709.41	0.005
20	ENSG00000166963	MAP1A	29	3.573477	1302.94	0.005
21	ENSG00000065559	MAP2K4	28	0.746543	1001.479	0.04
22	ENSG00000142208	AKT1	27	0.514225	5081.257	0.005236
23	ENSG00000185010	F8	27	-0.59579	-1480.95	0.02
24	ENSG00000042832	TG	26	4.192094	70.0546	0.005
25	ENSG00000149311	ATM	25	4.562618	2396.868	0.005

Continued on next page

Table B.9: **Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA (descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
26	ENSG00000101680	LAMA1	24	4.36995	330.5575	0.005
27	ENSG00000127329	PTPRB	24	1.833719	1538.255	0.005
28	ENSG00000178209	PLEC	24	0.554149	9368.297	0.005076
29	ENSG00000038427	VCAN	23	2.048457	16525.61	0.005
30	ENSG00000134982	APC	23	1.559689	1346.885	0.020942
31	ENSG00000065534	MYLK	23	1.243209	4132.391	0.005208
32	ENSG00000118058	KMT2A	23	0.889178	2384.59	0.005
33	ENSG00000142798	HSPG2	22	14.47531	17764.88	0.005
34	ENSG00000100503	NIN	22	1.540668	2161.873	0.005
35	ENSG00000130396	AFDN	22	0.995157	3900.523	0.005236
36	ENSG00000096433	ITPR3	22	0.843371	2741.744	0.005155
37	ENSG00000196367	TRRAP	22	0.759431	2868.675	0.005
38	ENSG00000167522	ANKRD11	22	0.504255	1578.139	0.005155
39	ENSG00000137076	TLN1	22	0.406877	7075.517	0.02
40	ENSG00000079102	RUNX1T1	22	-0.84228	-1998.55	0.015152
41	ENSG00000169946	ZFPM2	21	17.5438	389.4492	0.005
42	ENSG00000159216	RUNX1	21	12.13214	10676.8	0.02
43	ENSG00000134250	NOTCH2	21	2.218413	10127.18	0.005236
44	ENSG00000007314	SCN4A	21	1.620168	78.69703	0.005
45	ENSG00000029534	ANK1	20	32.53428	452.8636	0.005236
46	ENSG00000001626	CFTR	20	10.93548	57.16355	0.005
47	ENSG00000225830	ERCC6	20	4.851821	756.8204	0.031414
48	ENSG00000079841	RIMS1	20	1.219106	107.5541	0.005
49	ENSG00000110799	VWF	20	0.758277	4968.041	0.005076
50	ENSG00000070961	ATP2B1	20	0.508201	1021.101	0.005
51	ENSG00000139687	RB1	20	0.49751	1475.884	0.005076
52	ENSG00000138162	TACC2	20	0.41903	1064.017	0.005
53	ENSG00000112769	LAMA4	20	0.41313	1759.742	0.03
54	ENSG00000198646	NCOA6	20	0.28724	1181.123	0.005236
55	ENSG00000142949	PTPRF	20	-0.43415	-34519.6	0.025773
56	ENSG00000176406	RIMS2	19	67.2837	240.5586	0.005
57	ENSG00000197565	COL4A6	19	9.16545	594.5745	0.03
58	ENSG00000273079	GRIN2B	19	3.78143	18.91918	0.02
59	ENSG00000107186	MPDZ	19	0.93998	1025.01	0.005

Continued on next page

Table B.9: **Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA (descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
60	ENSG00000188153	COL4A5	19	0.58	1303.352	0.015
61	ENSG00000137474	MYO7A	19	0.31211	134.4706	0.03
62	ENSG00000082898	XPO1	19	0.14889	1508.017	0.005236
63	ENSG00000128591	FLNC	19	-0.48673	-1134.23	0.020725
64	ENSG00000142449	FBN3	19	-1.54672	589.6525	0.02
65	ENSG00000161681	SHANK1	19	-3.55523	80.60168	0.020942
66	ENSG00000081052	COL4A4	18	30.1425	181.2441	0.005
67	ENSG00000067369	TP53BP1	18	1.91657	2220.119	0.005
68	ENSG00000131626	PPFIA1	18	0.94291	2609.534	0.005
69	ENSG00000012048	BRCA1	18	0.59592	461.7295	0.045685
70	ENSG00000094631	HDAC6	18	0.47688	1059.178	0.005
71	ENSG00000147162	OGT	18	-0.39422	-6080.94	0.020942
72	ENSG00000146648	EGFR	17	6.012	4279.453	0.005236
73	ENSG00000147255	IGSF1	17	4.36987	1519.285	0.005236
74	ENSG00000163629	PTPN13	17	2.86052	2126.187	0.005
75	ENSG00000106278	PTPRZ1	17	2.689	210.4911	0.020942
76	ENSG00000018236	CNTN1	17	2.47952	566.872	0.045
77	ENSG00000169031	COL4A3	17	2.3773	50.87343	0.005
78	ENSG00000174469	CNTNAP2	17	1.27211	1129.258	0.005
79	ENSG00000100393	EP300	17	1.22172	4093.092	0.005155
80	ENSG00000185008	ROBO2	17	0.65907	281.5608	0.040404
81	ENSG00000102974	CTCF	17	0.62882	1733.975	0.005
82	ENSG00000112282	MED23	17	0.39046	1095.503	0.005
83	ENSG00000153201	RANBP2	17	0.29958	2693.065	0.005
84	ENSG00000101868	POLA1	17	0.20408	263.3489	0.020619
85	ENSG00000197081	IGF2R	17	-0.57152	-13446.8	0.005155
86	ENSG00000170776	AKAP13	17	-76.2496	8725.715	0.020942
87	ENSG00000198932	GPRASP1	16	211.617	435.8121	0.005236
88	ENSG00000077522	ACTN2	16	3.00835	100.4633	0.005
89	ENSG00000164692	COL1A2	16	1.73385	174362.1	0.005
90	ENSG00000089250	NOS1	16	1.45626	11.26545	0.005
91	ENSG00000198753	PLXNB3	16	1.02051	752.7958	0.010309
92	ENSG00000136869	TLR4	16	0.93304	475.4087	0.005
93	ENSG00000145675	PIK3R1	16	0.71824	3417.801	0.005

Continued on next page

Table B.9: **Top 100 dynamic cancer drivers from VIMtime(Bulk), ranked by frequency of mutation in TCGA (descending order).** (Continued)

	Ensembl.ID	HGNC.symbol	Mut.Freq	RelEffect	AbsEffect	p.val
94	ENSG00000141367	CLTC	16	0.71262	17864.83	0.005
95	ENSG00000197694	SPTAN1	16	0.58407	8909.812	0.015
96	ENSG00000114127	XRN1	16	0.51134	914.64	0.005
97	ENSG00000167306	MYO5B	16	0.37732	1542.141	0.020942
98	ENSG00000139618	BRCA2	16	0.29356	212.2813	0.03
99	ENSG00000085511	MAP3K4	16	0.22794	358.4106	0.031414
100	ENSG00000263528	IKBKE	16	-0.21876	-329.815	0.04712

B.12 Regulatory relationships among Dynamic Cancer Drivers - HER2time(SC)

Regulatory relationships between the top 200 of DCDs and the set (not restricted to the top 200 DCDs) of CGC Breast Cancer drivers detected by our method (when using "HER2time(SC)"). BRCA-TCGA Regulatory relationships extracted from http://www.grndb.com/

Table B.10: **Regulatory relationships among Dynamic Cancer Drivers - HER2time(SC).**

	TF	TF.isCGC	gene	gene.isCGC	Confidence
1	BCL3	FALSE	BCL3	FALSE	High
2	BCL3	FALSE	GNB2	FALSE	High
3	BCL3	FALSE	PRKACA	FALSE	High
4	BCL3	FALSE	RGS16	FALSE	High
5	BRCA1	TRUE	MED1	FALSE	High
6	BRCA1	TRUE	RFC5	FALSE	High
7	BRCA1	TRUE	BRCA1	TRUE	High
8	BRCA1	TRUE	MELK	FALSE	High
9	BRCA1	TRUE	TTF2	FALSE	High
10	BRCA1	TRUE	E2F2	FALSE	High
11	BRCA1	TRUE	CCND1	TRUE	High
12	BRCA1	TRUE	CDC25B	FALSE	High
13	BRCA1	TRUE	HTT	FALSE	High
14	BRCA1	TRUE	MYBL2	FALSE	High

Continued on next page

Table B.10: **Regulatory relationships among Dynamic Cancer Drivers - HER2time(SC).** (Continued)

	TF	TF.isCGC	gene	gene.isCGC	Confidence
15	BRCA1	TRUE	STAT5B	FALSE	High
16	BRCA1	TRUE	BRCA2	TRUE	High
17	BRCA1	TRUE	MCM6	FALSE	High
18	BRCA1	TRUE	TRIM37	FALSE	High
19	BRCA1	TRUE	CENPJ	FALSE	High
20	BRCA1	TRUE	CHEK2	TRUE	High
21	BRCA1	TRUE	EXOSC2	FALSE	High
22	BRCA1	TRUE	PIP4K2C	FALSE	High
23	BRCA1	TRUE	PLK1	FALSE	High
24	CTCF	TRUE	CDH1	TRUE	High
25	CTCF	TRUE	CSNK2A2	FALSE	High
26	CTCF	TRUE	CTCF	TRUE	High
27	E2F2	FALSE	CDC25B	FALSE	High
28	E2F2	FALSE	CENPJ	FALSE	High
29	E2F2	FALSE	MDC1	FALSE	High
30	E2F2	FALSE	PLK1	FALSE	High
31	E2F2	FALSE	CCT7	FALSE	High
32	E2F2	FALSE	CDKN2A	FALSE	High
33	E2F2	FALSE	MPHOSPH6	FALSE	High
34	E2F2	FALSE	TAF4	FALSE	High
35	E2F2	FALSE	TRIM37	FALSE	High
36	E2F2	FALSE	PTPN2	FALSE	High
37	E2F2	FALSE	BAG6	FALSE	High
38	E2F2	FALSE	PLCG1	FALSE	High
39	E2F2	FALSE	PIK3R2	FALSE	High
40	E2F2	FALSE	LSM2	FALSE	High
41	E2F2	FALSE	E2F2	FALSE	High
42	E2F2	FALSE	CDK1	FALSE	High
43	E2F2	FALSE	FEN1	TRUE	High
44	E2F2	FALSE	MCM6	FALSE	High
45	E2F2	FALSE	MYBL2	FALSE	High
46	E2F2	FALSE	SF3B3	FALSE	High
47	E2F2	FALSE	BRCA2	TRUE	High
48	E2F2	FALSE	CHEK2	TRUE	High

Continued on next page

Table B.10: **Regulatory relationships among Dynamic Cancer Drivers - HER2time(SC).** (Continued)

	TF	TF.isCGC	gene	gene.isCGC	Confidence
49	E2F2	FALSE	HTATSF1	FALSE	High
50	E2F2	FALSE	MELK	FALSE	High
51	E2F2	FALSE	BRCA1	TRUE	High
52	E2F2	FALSE	EXOSC2	FALSE	High
53	E2F2	FALSE	MTA2	FALSE	High
54	E2F2	FALSE	RFC5	FALSE	High
55	E2F2	FALSE	TTF2	FALSE	High
56	ESR1	TRUE	ERBB4	FALSE	High
57	ESR1	TRUE	NRIP1	FALSE	High
58	ESR1	TRUE	ESR1	TRUE	High
59	ESR1	TRUE	GAK	FALSE	High
60	ESR1	TRUE	IKBKB	FALSE	High
61	ESR1	TRUE	ERBB2	TRUE	High
62	HOXB7	FALSE	HOXB7	FALSE	High
63	MYBL2	FALSE	CENPJ	FALSE	High
64	MYBL2	FALSE	EXOSC2	FALSE	High
65	MYBL2	FALSE	MTA2	FALSE	High
66	MYBL2	FALSE	RFC5	FALSE	High
67	NFATC2	FALSE	CYFIP2	FALSE	High
68	NFATC2	FALSE	IL16	FALSE	High
69	NFATC2	FALSE	ITK	FALSE	High
70	RB1	TRUE	RB1	TRUE	High
71	SOX9	FALSE	REPS1	FALSE	High
72	SOX9	FALSE	SOX9	FALSE	High
73	STAT5B	FALSE	STAT5B	FALSE	High
74	THRA	FALSE	ERBB2	TRUE	High
75	THRA	FALSE	GRB7	FALSE	High
76	THRA	FALSE	THRA	FALSE	High
77	THRA	FALSE	STAT5B	FALSE	High
78	THRA	FALSE	NFE2L1	FALSE	High
79	THRA	FALSE	PLXNB1	FALSE	High
80	THRA	FALSE	UNC119	FALSE	High
81	THRA	FALSE	PKN1	FALSE	High

B.13 Regulatory relationships among Dynamic Cancer Drivers - HER2time(Bulk)

Regulatory relationships between the top 200 of DCDs and the set (not restricted to the top 200 DCDs) of CGC Breast Cancer drivers detected by our method (when using "HER2time(Bulk)"). BRCA-TCGA Regulatory relationships extracted from http://www.grndb.com/

Table B.11: **Regulatory relationships among Dynamic Cancer Drivers - HER2time(Bulk).**

	TF	TF.isCGC	gene	gene.isCGC	Confidence
1	BRCA1	TRUE	BRCA1	TRUE	High
2	BRCA1	TRUE	CCND1	TRUE	High
3	BRCA1	TRUE	CELSR3	FALSE	High
4	BRCA1	TRUE	HTT	FALSE	High
5	BRCA1	TRUE	SMARCC1	FALSE	High
6	BRCA1	TRUE	BARD1	TRUE	High
7	BRCA1	TRUE	PARP1	FALSE	High
8	BRCA1	TRUE	RFC1	FALSE	High
9	CTCF	TRUE	CDH1	TRUE	High
10	CTCF	TRUE	CTCF	TRUE	High
11	EP300	TRUE	EP300	TRUE	High
12	ESR1	TRUE	ITPR1	FALSE	High
13	ESR1	TRUE	WWP1	FALSE	High
14	ESR1	TRUE	ERBB4	FALSE	High
15	ESR1	TRUE	NEDD4L	FALSE	High
16	ESR1	TRUE	ESR1	TRUE	High
17	ESR1	TRUE	SPTBN4	FALSE	High
18	ESR1	TRUE	IKBKB	FALSE	High
19	ESR1	TRUE	GATA3	TRUE	High
20	ESR1	TRUE	ERBB2	TRUE	High
21	ESR1	TRUE	ITGA3	FALSE	High
22	ESR1	TRUE	FGFR2	FALSE	High
23	ESRRG	FALSE	ESRRG	FALSE	High
24	ESRRG	FALSE	SORBS2	FALSE	High
25	ESRRG	FALSE	CNTN2	FALSE	High
26	GATA3	TRUE	GATA3	TRUE	High

Continued on next page

Table B.11: **Regulatory relationships among Dynamic Cancer Drivers - HER2time(Bulk).** (Continued)

	TF	TF.isCGC	gene	gene.isCGC	Confidence
27	GATA3	TRUE	ESR1	TRUE	High
28	KMT2A	FALSE	ATM	FALSE	High
29	KMT2A	FALSE	EP300	TRUE	High
30	KMT2A	FALSE	SMARCA2	FALSE	High
31	KMT2A	FALSE	XRN1	FALSE	High
32	RFX3	FALSE	FGFR2	FALSE	High
33	RFX3	FALSE	MPDZ	FALSE	High
34	RFX3	FALSE	RFX3	FALSE	High
35	RFX3	FALSE	SMARCA2	FALSE	High
36	RFX3	FALSE	MAGI1	FALSE	High
37	RFX3	FALSE	CDKN1B	TRUE	High
38	RFX3	FALSE	ATP2B4	FALSE	High
39	RFX3	FALSE	NTRK2	FALSE	High
40	RFX3	FALSE	SORBS2	FALSE	High
41	RFX3	FALSE	COL4A5	FALSE	High
42	RFX3	FALSE	GLI3	FALSE	High
43	RUNX1	FALSE	RUNX1	FALSE	High
44	SIN3A	FALSE	SIN3A	FALSE	High
45	SIN3A	FALSE	DOCK1	FALSE	High
46	SIN3A	FALSE	IGF1R	FALSE	High
47	SIN3A	FALSE	CHD4	FALSE	High
48	SIN3A	FALSE	MAGI3	FALSE	High
49	SIN3A	FALSE	PER2	FALSE	High
50	SIN3A	FALSE	RFC1	FALSE	High
51	STAT5A	FALSE	STAT5A	FALSE	High
52	TP63	FALSE	COL17A1	FALSE	High
53	TP63	FALSE	DST	FALSE	High
54	TP63	FALSE	TP63	FALSE	High
55	TP63	FALSE	COL4A6	FALSE	High
56	TP63	FALSE	GRIA4	FALSE	High

B.14 Not differentially expressed DCD inferred from HER2Time(Bulk)

List with the DCDs inferred by using "HER2time(Bulk)" that are in the top 100 (ranked by frequency of mutation in TCGA (descending order)) and were detected as not differentially expressed between the conditions normal/cancer. We have included 2 flags indicating whether the DCD gene is a CGC gene/ a CGC Breast Cancer Driver.

Table B.12: **DCD inferred from HER2Time(Bulk) that are not differentially expressed**

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
1	ENSG00000121879	PIK3CA	TRUE	TRUE
2	ENSG00000039068	CDH1	TRUE	TRUE
3	ENSG00000141027	NCOR1	TRUE	TRUE
4	ENSG00000127603	MACF1	FALSE	FALSE
5	ENSG00000081479	LRP2	FALSE	FALSE
6	ENSG00000085224	ATRX	TRUE	FALSE
7	ENSG00000196712	NF1	TRUE	FALSE
8	ENSG00000166147	FBN1	FALSE	FALSE
9	ENSG00000171862	PTEN	TRUE	FALSE
10	ENSG00000111642	CHD4	TRUE	FALSE
11	ENSG00000065559	MAP2K4	TRUE	TRUE
12	ENSG00000142208	AKT1	TRUE	TRUE
13	ENSG00000104517	UBR5	TRUE	FALSE
14	ENSG00000149311	ATM	TRUE	FALSE
15	ENSG00000178209	PLEC	FALSE	FALSE
16	ENSG00000178104	PDE4DIP	TRUE	FALSE
17	ENSG00000118058	KMT2A	TRUE	FALSE
18	ENSG00000130396	AFDN	TRUE	FALSE
19	ENSG00000278540	ACACA	FALSE	FALSE
20	ENSG00000142798	HSPG2	FALSE	FALSE
21	ENSG00000137076	TLN1	FALSE	FALSE
22	ENSG00000167522	ANKRD11	FALSE	FALSE
23	ENSG00000196367	TRRAP	TRUE	FALSE
24	ENSG00000159216	RUNX1	TRUE	FALSE
25	ENSG00000134250	NOTCH2	TRUE	FALSE

Continued on next page

Table B.12: **DCD inferred from HER2Time(Bulk) that are not differentially expressed** (Continued)

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
26	ENSG00000029534	ANK1	TRUE	FALSE
27	ENSG00000001626	CFTR	FALSE	FALSE
28	ENSG00000138162	TACC2	FALSE	FALSE
29	ENSG00000142949	PTPRF	FALSE	FALSE
30	ENSG00000197386	HTT	FALSE	FALSE
31	ENSG00000225830	ERCC6	FALSE	FALSE
32	ENSG00000150995	ITPR1	FALSE	FALSE
33	ENSG00000188153	COL4A5	FALSE	FALSE
34	ENSG00000081052	COL4A4	FALSE	FALSE
35	ENSG00000064012	CASP8	TRUE	TRUE
36	ENSG00000131626	PPFIA1	FALSE	FALSE
37	ENSG00000147255	IGSF1	FALSE	FALSE
38	ENSG00000018236	CNTN1	FALSE	FALSE
39	ENSG00000112282	MED23	FALSE	FALSE
40	ENSG00000170776	AKAP13	FALSE	FALSE
41	ENSG00000197081	IGF2R	FALSE	FALSE
42	ENSG00000100393	EP300	TRUE	TRUE
43	ENSG00000102974	CTCF	TRUE	TRUE
44	ENSG00000132694	ARHGEF11	FALSE	FALSE
45	ENSG00000153201	RANBP2	TRUE	FALSE
46	ENSG00000178568	ERBB4	TRUE	FALSE
47	ENSG00000198753	PLXNB3	FALSE	FALSE
48	ENSG00000167306	MYO5B	FALSE	FALSE
49	ENSG00000197694	SPTAN1	FALSE	FALSE
50	ENSG00000114127	XRN1	FALSE	FALSE
51	ENSG00000049759	NEDD4L	FALSE	FALSE
52	ENSG00000196313	POM121	FALSE	FALSE
53	ENSG00000169855	ROBO1	FALSE	FALSE
54	ENSG00000183454	GRIN2A	TRUE	FALSE
55	ENSG00000155511	GRIA1	FALSE	FALSE
56	ENSG00000132849	PATJ	FALSE	FALSE
57	ENSG00000151276	MAGI1	FALSE	FALSE
58	ENSG00000130985	UBA1	FALSE	FALSE
59	ENSG00000196083	IL1RAP	FALSE	FALSE

Continued on next page

Table B.12: **DCD inferred from HER2Time(Bulk) that are not differentially expressed** (Continued)

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
60	ENSG00000197535	MYO5A	TRUE	FALSE
61	ENSG00000182752	PAPPA	FALSE	FALSE
62	ENSG00000106571	GLI3	FALSE	FALSE
63	ENSG00000100852	ARHGAP5	TRUE	FALSE
64	ENSG00000124151	NCOA3	FALSE	FALSE
65	ENSG00000112851	ERBIN	FALSE	FALSE

B.15 Not differentially expressed DCD inferred from VIMTime(Bulk)

List with the DCDs inferred by using "VIMtime(Bulk)" that are in the top 100 (ranked by frequency of mutation in TCGA (descending order)) and were detected as not differentially expressed between the conditions normal/cancer. We have included 2 flags indicating whether the DCD gene is a CGC gene/ a CGC Breast Cancer Driver.

Table B.13: **DCD inferred from VIMTime(Bulk) that are not differentially expressed**

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
1	ENSG00000121879	PIK3CA	TRUE	TRUE
2	ENSG00000198626	RYR2	FALSE	FALSE
3	ENSG00000163554	SPTA1	FALSE	FALSE
4	ENSG00000039068	CDH1	TRUE	TRUE
5	ENSG00000127603	MACF1	FALSE	FALSE
6	ENSG00000081479	LRP2	FALSE	FALSE
7	ENSG00000085224	ATRX	TRUE	FALSE
8	ENSG00000127914	AKAP9	TRUE	FALSE
9	ENSG00000196712	NF1	TRUE	FALSE
10	ENSG00000065526	SPEN	TRUE	FALSE
11	ENSG00000166147	FBN1	FALSE	FALSE
12	ENSG00000111642	CHD4	TRUE	FALSE
13	ENSG00000166963	MAP1A	FALSE	FALSE
14	ENSG00000065559	MAP2K4	TRUE	TRUE

Continued on next page

Table B.13: **DCD inferred from VIMTime(Bulk) that are not differentially expressed** (Continued)

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
15	ENSG00000142208	AKT1	TRUE	TRUE
16	ENSG00000149311	ATM	TRUE	FALSE
17	ENSG00000178209	PLEC	FALSE	FALSE
18	ENSG00000134982	APC	TRUE	FALSE
19	ENSG00000118058	KMT2A	TRUE	FALSE
20	ENSG00000142798	HSPG2	FALSE	FALSE
21	ENSG00000100503	NIN	TRUE	FALSE
22	ENSG00000130396	AFDN	TRUE	FALSE
23	ENSG00000096433	ITPR3	FALSE	FALSE
24	ENSG00000196367	TRRAP	TRUE	FALSE
25	ENSG00000167522	ANKRD11	FALSE	FALSE
26	ENSG00000137076	TLN1	FALSE	FALSE
27	ENSG00000159216	RUNX1	TRUE	FALSE
28	ENSG00000134250	NOTCH2	TRUE	FALSE
29	ENSG00000029534	ANK1	TRUE	FALSE
30	ENSG00000001626	CFTR	FALSE	FALSE
31	ENSG00000225830	ERCC6	FALSE	FALSE
32	ENSG00000070961	ATP2B1	FALSE	FALSE
33	ENSG00000139687	RB1	TRUE	TRUE
34	ENSG00000138162	TACC2	FALSE	FALSE
35	ENSG00000198646	NCOA6	FALSE	FALSE
36	ENSG00000142949	PTPRF	FALSE	FALSE
37	ENSG00000188153	COL4A5	FALSE	FALSE
38	ENSG00000137474	MYO7A	FALSE	FALSE
39	ENSG00000082898	XPO1	TRUE	FALSE
40	ENSG00000081052	COL4A4	FALSE	FALSE
41	ENSG00000067369	TP53BP1	FALSE	FALSE
42	ENSG00000131626	PPFIA1	FALSE	FALSE
43	ENSG00000094631	HDAC6	FALSE	FALSE
44	ENSG00000147162	OGT	FALSE	FALSE
45	ENSG00000147255	IGSF1	FALSE	FALSE
46	ENSG00000163629	PTPN13	TRUE	FALSE
47	ENSG00000018236	CNTN1	FALSE	FALSE
48	ENSG00000169031	COL4A3	FALSE	FALSE

Continued on next page

Table B.13: **DCD inferred from VIMTime(Bulk) that are not differentially expressed** (Continued)

	Ensembl.ID	HGNC.symbol	is.CGC	is.BRCA_CGC
49	ENSG00000100393	EP300	TRUE	TRUE
50	ENSG00000102974	CTCF	TRUE	TRUE
51	ENSG00000112282	MED23	FALSE	FALSE
52	ENSG00000153201	RANBP2	TRUE	FALSE
53	ENSG00000101868	POLA1	FALSE	FALSE
54	ENSG00000197081	IGF2R	FALSE	FALSE
55	ENSG00000170776	AKAP13	FALSE	FALSE
56	ENSG00000198753	PLXNB3	FALSE	FALSE
57	ENSG00000197694	SPTAN1	FALSE	FALSE
58	ENSG00000114127	XRN1	FALSE	FALSE
59	ENSG00000167306	MYO5B	FALSE	FALSE
60	ENSG00000085511	MAP3K4	FALSE	FALSE

B.16 DCD inferred from single cell data when using Monocle3 pseudotime (event detection using HER2).

Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2), ranked by Relative Causal Impact (descending order). A full comprehensive list (687 DCDs) can be found in *"Supplementary/ supplementary table 17 - DCDs(HER2)monocle3_SC.csv"* at `github.com/AndresMCB/DynamicCancerDriver`. Alternatively, it can be accessed at `https://doi.org/10.6084/m9.figshare.20448384`

Table B.14: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2).**

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
1	ENSG00000026025	VIM	5497.015677	48.46322256	0.005
2	ENSG00000150907	FOXO1	4714.552873	5.608519119	0.005
3	ENSG00000131981	LGALS3	4144.929251	84.26206587	0.0053191
4	ENSG00000090104	RGS1	2903.345228	87.32849747	0.0054644
5	ENSG00000146648	EGFR	2355.218769	6.286198263	0.005
6	ENSG00000007372	PAX6	2125.433108	9.776856375	0.0054644

Continued on next page

Table B.14: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
7	ENSG00000139618	BRCA2	1929.235161	10.1325149	0.0054644
8	ENSG00000163629	PTPN13	1432.959532	9.865542452	0.0053191
9	ENSG00000113580	NR3C1	1331.273657	18.72189155	0.0054644
10	ENSG00000170312	CDK1	1324.514951	38.18243465	0.005
11	ENSG00000145675	PIK3R1	1268.243844	43.10251887	0.005
12	ENSG00000204388	HSPA1B	875.1034225	82.62241402	0.0054644
13	ENSG00000147168	IL2RG	873.7077514	24.32384305	0.0054644
14	ENSG00000115170	ACVR1	866.3055373	11.72843566	0.0054644
15	ENSG00000065534	MYLK	863.2676005	19.83828168	0.0054644
16	ENSG00000116478	HDAC1	849.4868708	20.04177936	0.005
17	ENSG00000008710	PKD1	678.7199206	1.631221216	0.005
18	ENSG00000057657	PRDM1	645.531052	8.050021523	0.005
19	ENSG00000134954	ETS1	385.4317008	3.463086243	0.005
20	ENSG00000161638	ITGA5	364.1595487	9.65155461	0.0052356
21	ENSG00000105810	CDK6	327.1609314	10.69510869	0.005
22	ENSG00000211899	IGHM	296.2099599	5.988681984	0.005
23	ENSG00000162511	LAPTM5	295.1172371	14.75063261	0.0054644
24	ENSG00000120899	PTK2B	283.5072807	5.138283775	0.005
25	ENSG00000130699	TAF4	270.6514607	3.524711563	0.005
26	ENSG00000061273	HDAC7	256.0360489	7.112382978	0.0054644
27	ENSG00000171425	ZNF581	250.9541922	10.09179279	0.005
28	ENSG00000165392	WRN	228.0468516	9.058240204	0.0052356
29	ENSG00000071539	TRIP13	223.1553307	4.448673093	0.0054644
30	ENSG00000169554	ZEB2	210.4185352	7.532648196	0.005
31	ENSG00000146232	NFKBIE	201.0148845	8.677575516	0.0054644
32	ENSG00000182195	LDOC1	194.4252793	34.65212748	0.005
33	ENSG00000086015	MAST2	186.2171539	4.115568991	0.0053191
34	ENSG00000157933	SKI	185.711374	1.03574966	0.0053191
35	ENSG00000080839	RBL1	160.4416156	5.529130287	0.0054644
36	ENSG00000163565	IFI16	136.502815	24.41421937	0.005
37	ENSG00000100297	MCM5	123.7222159	8.307642107	0.0054644
38	ENSG00000113140	SPARC	121.0313529	13.42570679	0.005
39	ENSG00000103168	TAF1C	109.1031238	2.267694096	0.005
40	ENSG00000100365	NCF4	103.6331154	2.048389558	0.0054644

Continued on next page

Table B.14: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
41	ENSG00000110031	LPXN	101.9555336	10.79265484	0.005
42	ENSG00000123609	NMI	100.1228912	17.83350199	0.005
43	ENSG00000079819	EPB41L2	97.43900444	13.96368337	0.0054644
44	ENSG00000094804	CDC6	95.56694704	54.24212624	0.0054644
45	ENSG00000181191	PJA1	92.19887376	10.00677279	0.0053191
46	ENSG00000115904	SOS1	87.69865463	19.24214135	0.0052356
47	ENSG00000204389	HSPA1A	76.13635005	251.2691652	0.0054644
48	ENSG00000107438	PDLIM1	72.05086323	51.4028857	0.005
49	ENSG00000163602	RYBP	67.03034468	8.669466149	0.005
50	ENSG00000153094	BCL2L11	59.43233576	18.57503198	0.005
51	ENSG00000112242	E2F3	59.01830337	5.408361	0.0054644
52	ENSG00000076003	MCM6	54.98556671	11.11409	0.0053191
53	ENSG00000169679	BUB1	54.42728252	5.304895	0.005
54	ENSG00000011243	AKAP8L	47.46843703	36.62557	0.0052356
55	ENSG00000149269	PAK1	46.11539601	30.73035	0.0053191
56	ENSG00000077235	GTF3C1	43.64561452	12.07654	0.0054644
57	ENSG00000150995	ITPR1	42.72549849	11.15117	0.005
58	ENSG00000076984	MAP2K7	40.18901772	0.720771	0.005
59	ENSG00000158716	DUSP23	37.49162824	20.19945	0.0054644
60	ENSG00000118785	SPP1	36.6437301	60.92617	0.0054644
61	ENSG00000130066	SAT1	34.5814983	8263.452	0.0054644
62	ENSG00000185591	SP1	33.33557001	13.55126	0.005
63	ENSG00000115414	FN1	31.9640872	53.38424	0.005
64	ENSG00000126351	THRA	30.44203132	7.15252	0.0054644
65	ENSG00000175899	A2M	30.00699549	128.5976	0.005
66	ENSG00000101224	CDC25B	29.95894139	8.40489	0.005
67	ENSG00000131323	TRAF3	29.22687329	2.972127	0.005
68	ENSG00000101868	POLA1	26.83742502	4.764538	0.0054644
69	ENSG00000089685	BIRC5	26.67220914	20.99904	0.0218579
70	ENSG00000106799	TGFBR1	25.68393348	19.54363	0.005
71	ENSG00000125817	CENPB	25.27345499	0.877761	0.0053191
72	ENSG00000173020	GRK2	25.05807981	1.365911	0.005
73	ENSG00000111445	RFC5	24.7227333	14.36336	0.005
74	ENSG00000173757	STAT5B	23.72027886	6.133278	0.005

Continued on next page

Table B.14: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using HER2).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
75	ENSG00000117394	SLC2A1	23.65214494	37.57094	0.0050251
76	ENSG00000116809	ZBTB17	23.09264935	16.5387	0.0053191
77	ENSG00000196482	ESRRG	21.97905737	12.7691	0.0054644
78	ENSG00000185499	MUC1	21.6742046	260.7841	0.005
79	ENSG00000129245	FXR2	21.58980148	6.144121	0.0053191
80	ENSG00000171223	JUNB	21.16142238	19.90677	0.005
81	ENSG00000169375	SIN3A	20.91266456	65.02031	0.005
82	ENSG00000072201	LNX1	20.77137243	13.89379	0.0052356
83	ENSG00000132002	DNAJB1	20.52972325	130.2722	0.0054644
84	ENSG00000072062	PRKACA	20.45013863	3.467959	0.005
85	ENSG00000143622	RIT1	19.37067322	61.68632	0.0054644
86	ENSG00000108262	GIT1	18.86727341	4.247711	0.0053191
87	ENSG00000168395	ING5	18.19410268	6.602726	0.005
88	ENSG00000149257	SERPINH1	17.99791167	6.785631	0.005
89	ENSG00000153487	ING1	17.61159174	5.329368	0.005
90	ENSG00000140564	FURIN	17.05784999	3.395597	0.005
91	ENSG00000141646	SMAD4	16.78768469	13.59412	0.005
92	ENSG00000067369	TP53BP1	16.07849707	33.25126	0.0052356
93	ENSG00000136527	TRA2B	15.85084817	221.5867	0.005
94	ENSG00000150760	DOCK1	15.81835072	7.598781	0.005
95	ENSG00000092820	EZR	15.30487631	643.9626	0.0218579
96	ENSG00000117632	STMN1	15.25582873	324.2975	0.0054644
97	ENSG00000111348	ARHGDIB	14.79380311	205.5064	0.0054644
98	ENSG00000101400	SNTA1	14.60203488	3.230965	0.0054644
99	ENSG00000134899	ERCC5	14.49894363	16.71281	0.0054644
100	ENSG00000105229	PIAS4	14.33159882	2.572477	0.0054644

B.17 DCD inferred from single cell data when using Monocle3 pseudotime(event detection using VIM).

Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM), ranked by Relative Causal Impact (descending order). A full comprehensive list (665 DCDs) can be found in *"Supplementary/ supplementary table 18 - DCDs(VIM)monocle3_SC.csv"* at github.com/AndresMCB/

DynamicCancerDriver. Alternatively, it can be accessed at https://doi.org/10.6084/m9.figshare.20448384

Table B.15: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM).**

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
1	ENSG00000095066	HOOK2	5167.37962	27.77482041	0.021857923
2	ENSG00000090104	RGS1	3643.090673	87.90355845	0.005464481
3	ENSG00000146648	EGFR	3112.550314	6.327804137	0.005
4	ENSG00000131981	LGALS3	3005.447105	84.80324928	0.005464481
5	ENSG00000139618	BRCA2	2429.572877	10.19961297	0.005464481
6	ENSG00000113580	NR3C1	1724.40904	18.84708344	0.005464481
7	ENSG00000145675	PIK3R1	1680.035878	43.39169659	0.005
8	ENSG00000116478	HDAC1	1124.17988	20.17814199	0.005
9	ENSG00000147168	IL2RG	1098.621453	24.48803595	0.005464481
10	ENSG00000115170	ACVR1	1049.61868	11.80722	0.005464481
11	ENSG00000118971	CCND2	987.3905813	5.573443329	0.005464481
12	ENSG00000057657	PRDM1	834.2949252	8.105301136	0.005
13	ENSG00000129675	ARHGEF6	821.780364	6.617159502	0.005464481
14	ENSG00000204388	HSPA1B	770.6368303	83.14673187	0.005464481
15	ENSG00000065534	MYLK	752.0981175	19.96368402	0.005464481
16	ENSG00000211899	IGHM	499.8789776	6.035970738	0.005
17	ENSG00000157110	RBPMS	419.6438232	19.04546208	0.005
18	ENSG00000105810	CDK6	371.5603146	10.76870491	0.005
19	ENSG00000134954	ETS1	340.6730401	3.483910257	0.005
20	ENSG00000130699	TAF4	329.8696983	3.550019874	0.005
21	ENSG00000162511	LAPTM5	318.33128	14.85038511	0.005464481
22	ENSG00000169554	ZEB2	305.5653887	7.59290374	0.005
23	ENSG00000071539	TRIP13	293.700486	4.482457866	0.005464481
24	ENSG00000123609	NMI	203.4667534	18.03225441	0.005
25	ENSG00000086015	MAST2	199.5809937	4.143862595	0.005464481
26	ENSG00000163565	IFI16	194.1566688	24.62645186	0.005
27	ENSG00000080839	RBL1	190.9648538	5.570665975	0.005464481
28	ENSG00000146232	NFKBIE	173.4138221	8.726712779	0.005464481
29	ENSG00000113140	SPARC	165.9340664	13.54320287	0.005
30	ENSG00000079819	EPB41L2	165.0299126	14.11337242	0.005464481
31	ENSG00000126561	STAT5A	158.1172205	7.212139828	0.005

Continued on next page

Table B.15: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
32	ENSG00000061273	HDAC7	156.6370247	7.138336323	0.005464481
33	ENSG00000100297	MCM5	155.7528143	8.375573751	0.005464481
34	ENSG00000100365	NCF4	135.1333586	2.066337565	0.005464481
35	ENSG00000094804	CDC6	131.2079871	54.74898538	0.005464481
36	ENSG00000175899	A2M	129.6239742	132.7250018	0.005
37	ENSG00000110031	LPXN	110.1277824	10.87080061	0.005
38	ENSG00000077235	GTF3C1	99.70110306	12.30827464	0.005464481
39	ENSG00000157933	SKI	98.25606218	1.037325369	0.005319149
40	ENSG00000204389	HSPA1A	97.73041504	253.6317525	0.005464481
41	ENSG00000181191	PJA1	96.41729107	10.07669394	0.005464481
42	ENSG00000107438	PDLIM1	94.04056497	51.90389966	0.005
43	ENSG00000076003	MCM6	73.13006448	11.23628687	0.005464481
44	ENSG00000169679	BUB1	72.33922917	5.36341446	0.005
45	ENSG00000120899	PTK2B	71.11948886	5.118036096	0.005
46	ENSG00000182195	LDOC1	64.67195828	34.5234399	0.005
47	ENSG00000141736	ERBB2	53.78312395	100.8353444	0.005
48	ENSG00000150995	ITPR1	53.25922334	11.27481766	0.005
49	ENSG00000076984	MAP2K7	43.35798709	0.726756156	0.005
50	ENSG00000136527	TRA2B	40.73948841	230.7600479	0.005
51	ENSG00000171425	ZNF581	37.92648	9.866243113	0.005
52	ENSG00000126351	THRA	34.58125	7.225962477	0.005464481
53	ENSG00000101868	POLA1	34.4132	4.833803455	0.005464481
54	ENSG00000106799	TGFBR1	34.32812	19.85816402	0.005
55	ENSG00000101224	CDC25B	33.83564	8.491069067	0.005
56	ENSG00000158092	NCK1	32.89718	37.94398278	0.005
57	ENSG00000149269	PAK1	32.82206	30.64050526	0.005235602
58	ENSG00000072201	LNX1	31.92938	14.21180431	0.005464481
59	ENSG00000008710	PKD1	30.76622	1.590865109	0.005
60	ENSG00000116809	ZBTB17	29.1999	16.79221854	0.005464481
61	ENSG00000175324	LSM1	28.95422	78.85814772	0.005
62	ENSG00000196482	ESRRG	27.85875	12.97142189	0.005464481
63	ENSG00000011243	AKAP8L	27.61783	36.30622058	0.005
64	ENSG00000153487	ING1	26.39598	5.45901815	0.005
65	ENSG00000089685	BIRC5	26.19338	21.1218908	0.021857923

Continued on next page

Table B.15: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
66	ENSG00000115904	SOS1	26.06123	18.86429799	0.005235602
67	ENSG00000158716	DUSP23	25.46537	20.06249137	0.005464481
68	ENSG00000185499	MUC1	24.50736	263.7965821	0.005
69	ENSG00000143622	RIT1	24.07303	62.68753871	0.005464481
70	ENSG00000131323	TRAF3	23.8326	2.968599191	0.005
71	ENSG00000175115	PACS1	23.41073	5.53003604	0.027322404
72	ENSG00000129245	FXR2	22.4708	6.192903743	0.005319149
73	ENSG00000118785	SPP1	20.9329	60.07447256	0.005464481
74	ENSG00000137337	MDC1	20.24821	7.172730382	0.005
75	ENSG00000132002	DNAJB1	20.11164	130.9474113	0.005464481
76	ENSG00000165392	WRN	19.85458	8.712394816	0.005
77	ENSG00000111276	CDKN1B	18.5527	7.665552551	0.005464481
78	ENSG00000111348	ARHGDIB	18.11962	209.2773074	0.005464481
79	ENSG00000149257	SERPINH1	17.90422	6.822094239	0.005
80	ENSG00000149554	CHEK1	17.80112	22.87938021	0.021857923
81	ENSG00000140939	NOL3	17.65267	11.17505869	0.005
82	ENSG00000150907	FOXO1	17.48704	5.332548906	0.005
83	ENSG00000125817	CENPB	17.30969	0.867197468	0.005319149
84	ENSG00000114107	CEP70	17.13509	34.15585737	0.02
85	ENSG00000177606	JUN	16.87409	24.06345225	0.005
86	ENSG00000140396	NCOA2	16.59145	10.73908573	0.005464481
87	ENSG00000067369	TP53BP1	15.95731	33.372974	0.005464481
88	ENSG00000132964	CDK8	15.8527	11.8504762	0.005464481
89	ENSG00000007372	PAX6	15.03364	9.211477627	0.021857923
90	ENSG00000163629	PTPN13	13.68312	9.239446515	0.005
91	ENSG00000198793	MTOR	13.51035	10.26927655	0.021857923
92	ENSG00000104765	BNIP3L	13.41124	108.5863932	0.005235602
93	ENSG00000155657	TTN	13.2287	1.290281094	0.005464481
94	ENSG00000135723	FHOD1	13.0542	12.15724464	0.005
95	ENSG00000147044	CASK	12.74202	45.51641543	0.005464481
96	ENSG00000124357	NAGK	12.05479	69.43220162	0.005
97	ENSG00000204469	PRRC2A	12.02893	5.142489858	0.005464481
98	ENSG00000115414	FN1	11.83336	50.99172744	0.005
99	ENSG00000169375	SIN3A	11.76968	63.0743067	0.005

Continued on next page

Table B.15: **Dynamic cancer drivers inferred from single cell data when using Monocle3 pseudotime (event detection using VIM).** (Continued)

	Ensembl.ID	HGNC.symbol	RelEffect	AbsEffect	p
100	ENSG00000091073	DTX2	11.60333	6.540285256	0.021857923

Appendix C

Supplementary material of identifying genes that cause cancer progression using dynamic causal inference

C.1 Hypergeometric test for the 39 analysed targets in Bulk Data

Hypergeometric test p.value was calculated by using the phyper function in R ($p.value = 1 - phyper(q-1, m, n, k)$). Parameters of the test are as follows: q = number of CGC genes among the inferred driver genes (column "CGC detected"), m = number of CGC genes among the model variables (column "CGC baseline"), n = number on non-CGC genes among the putative drivers (difference between columns "nVars" and "CGC baseline"), and k = number of inferred drivers.

Values in column "p.val 50" are calculated in a similar way, by setting q as "CGC in top150", and k as 150.

Results of our experiments have a statistically significant overlap with the CGC gene set for 11 target genes when considering the full list of inferred drivers. Furthermore, we have found a significant overlap with the CGC set for seven additional targets when using the *driver score* as the metric to select the top 150 inferred genes.

Table C.1: **Results of hypergeometric test for the 39 analysed targets in the TCGA-BRCA dataset.**

	Target y	nVars	CGC baseline	Inferred drivers	CGC detected	p.val	CGC in top150	p.val top150
1	SMAD4	733	94	350	50	0.1537	27	0.0262
2	USP9X	733	93	304	37	0.6781	17	0.7535
3	FOXP1	733	94	285	44	0.0585	26	0.0462
4	MEN1	733	94	351	56	0.0101	33	0.0003
5	AFDN	733	94	356	54	0.0414	21	0.3581
6	TBL1XR1	733	94	318	47	0.1015	24	0.1227
7	ERBB2	732	93	361	47	0.4438	21	0.3397
8	PIK3R1	732	93	445	69	0.0028	29	0.0061
9	KDM6A	733	94	280	41	0.1483	24	0.1227
10	BRCA1	732	93	361	50	0.2099	20	0.4437
11	TP53	733	94	277	44	0.0355	26	0.0462
12	ARID1A	733	94	345	49	0.1730	23	0.1846
13	NF1	733	94	317	50	0.0246	23	0.1846
14	SF3B1	733	94	342	47	0.2790	25	0.0773

Continued on next page

Table C.1: **Results of hypergeometric test for the 39 analysed targets in the TCGA-BRCA dataset.** (Continued)

	Target y	nVars	CGC baseline	Inferred drivers	CGC detected	p.val	CGC in top150	p.val top150
15	MAP2K4	733	94	297	40	0.3736	17	0.7703
16	RB1	733	94	323	45	0.2462	21	0.3581
17	CDH1	733	94	421	61	0.0721	30	0.0034
18	CDKN1B	733	94	280	42	0.1024	20	0.4632
19	PIK3CA	733	94	351	53	0.0489	22	0.2637
20	KMT2C	733	94	262	38	0.1839	30	0.0034
21	FOXO3	733	94	367	56	0.0309	24	0.1227
22	PTEN	733	94	369	50	0.3152	23	0.1846
23	BRCA2	732	93	327	53	0.0074	25	0.0700
24	AKT1	733	94	445	61	0.2195	18	0.6770
25	TBX3	733	94	417	56	0.3270	19	0.5722
26	CHEK2	733	94	278	45	0.0228	26	0.0462
27	CBFB	733	94	282	44	0.0488	22	0.2637
28	NCOR1	733	94	297	38	0.5505	27	0.0262
29	KRAS	733	94	321	40	0.6434	18	0.6770
30	CDKN2A	732	93	198	38	0.0014	31	0.0013
31	RUNX1	733	94	356	50	0.1976	23	0.1846
32	ZFP36L1	733	93	385	55	0.1043	25	0.0691
33	GATA3	732	93	361	39	0.9492	17	0.7557
34	MAP3K1	733	94	354	53	0.0582	25	0.0773
35	GPS2	733	93	459	58	0.5697	22	0.2452
36	CTCF	733	94	364	52	0.1434	23	0.1846
37	CTNNA1	733	93	361	51	0.1485	29	0.0060
38	BAP1	733	94	421	54	0.5451	21	0.3581
39	PBRM1	733	94	364	52	0.1434	27	0.0262

Appendix D

Statements of authorship

D.1 A pseudotemporal causality approach to indentifying miRNA-mRNA interactions during biological processess

University of South Australia

Statement of Authorship

Title of Paper	A pseudotemporal causality approach to identifying miRNA–mRNA interactions during biological processes
Publication Status	■ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details Relevant reference details	Andres M Cifuentes-Bernal, Vu Vh Pham, Xiaomei Li, Lin Liu, Jiuyong Li, Thuc Duy Le, A pseudotemporal causality approach to identifying miRNA–mRNA interactions during biological processes, Bioinformatics, Volume 37, Issue 6, March 2021, Pages 807–814, https://doi.org/10.1093/bioinformatics/btaa899
PRINCIPAL AUTHOR	
Name of Principal Author (Candidate)	Andres Mauricio Cifuentes Bernal
Contribution to the Paper Brief description of your work in this publication	Andres Cifuentes (A.C.). and Thuc Le (T.L.). conceived the experiments, A.C., Vu V H Pham (V.P). and Xiaomei Li (X.L.) conducted the experiments. A.C. and T.L analysed the results. A.C. and T.L. wrote the manuscript. V.P., X.L, Lin Liu (L.L.) and Jiuyong Li (J.L.) reviewed the manuscript.
Overall Percentage (%)	60%
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature *Signatures will be redacted by the Library at the time of publication*	Andres Cifuentes Digitally signed by Andres Cifuentes Date: 2023.11.22 13:58:32 +10'30'
Date	22-Nov-23
CO-AUTHOR CONTRIBUTIONS	

By signing the Statement of Authorship, each author certifies that:

I. the candidate's stated contribution to the publication is accurate (as detailed above);

II. permission is granted for the candidate in include the publication in the thesis; and

III. the sum of all co-author contribution is equal to 100% less the candidate's stated contribution.

Name of Co-Author 1	Vu Vh Pham

University of South Australia

% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Hoang	Date 24 Nov 23
Name of Co-Author 2	Xiaomei Li	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Xiaomei Li Digitally signed by Xiaomei Li Date: 2023.11.23 09:17:58 +11'00'	Date 23/11/2023
Name of Co-Author 3	Lin Liu	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Lin Liu Digitally signed by Lin Liu Date: 2023.11.22 20:55:42 +10'30'	Date 22/11/2023
Name of Co-Author 4	Jiuyong Li	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*		Date 27/11/2023
Name of Co-Author 5	Thuc Duy Le	
% Contribution to the Paper	16%	
Signature *Signatures will be redacted by the Library at the time of publication*		Date 22/11/23
Name of Co-Author 6		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 7		
% Contribution to the Paper		

Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 8		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 9		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 10		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date

D.2 Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories

University of South Australia

Statement of Authorship

Title of Paper	Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories
Publication Status	■ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details Relevant reference details	Andres M Cifuentes-Bernal, Vu V H Pham, Xiaomei Li, Lin Liu, Jiuyong Li, Thuc Duy Le, Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories, Briefings in Functional Genomics, Volume 21, Issue 6, November 2022, Pages 455–465, https://doi.org/10.1093/bfgp/elac030
PRINCIPAL AUTHOR	
Name of Principal Author (Candidate)	Andres Mauricio Cifuentes Bernal
Contribution to the Paper Brief description of your work in this publication	Andres Cifuentes (A.C.). and Thuc Le (T.L.). conceived the experiments, A.C., Vu V H Pham (V.P). and Xiaomei Li (X.L.) conducted the experiments. A.C. and T.L analysed the results. A.C. and T.L. wrote the manuscript. V.P., X.L, Lin Liu (L.L.) and Jiuyong Li (J.L.) reviewed the manuscript.
Overall Percentage (%)	60
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature *Signatures will be redacted by the Library at the time of publication*	Andres Cifuentes Digitally signed by Andres Cifuentes Date: 2023.11.22 13:59:09 +10'30'
Date	22-Nov-23
CO-AUTHOR CONTRIBUTIONS	

By signing the Statement of Authorship, each author certifies that:

I. the candidate's stated contribution to the publication is accurate (as detailed above);

II. permission is granted for the candidate in include the publication in the thesis; and

III. the sum of all co-author contribution is equal to 100% less the candidate's stated contribution.

Name of Co-Author 1	Vu Vh Pham

University of South Australia

% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Hoang	Date 24 Nov 23
Name of Co-Author 2	Xiaomei Li	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Xiaomei Li Digitally signed by Xiaomei Li Date: 2023.11.23 09:18:37 +11'00'	Date 23/11/2023
Name of Co-Author 3	Lin Liu	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*	Lin Liu Digitally signed by Lin Liu Date: 2023.11.22 20:54:42 +10'30'	Date 22/11/2023
Name of Co-Author 4	Jiuyong Li	
% Contribution to the Paper	6%	
Signature *Signatures will be redacted by the Library at the time of publication*		Date 27/11/2023
Name of Co-Author 5	Thuc Duy Le	
% Contribution to the Paper	16%	
Signature *Signatures will be redacted by the Library at the time of publication*		Date 22/11/23
Name of Co-Author 6		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 7		
% Contribution to the Paper		

University of South Australia

Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 8		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 9		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 10		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date

D.3 Identifying cooperative genes that cause cancer progression using dynamic causal inference

University of South Australia

Statement of Authorship

Title of Paper	Identifying cooperative genes that cause cancer progression using dynamic causal inference
Publication Status	☐ Published ☐ Accepted for Publication ■ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details Relevant reference details	Submitted to "Briefings in Functional Genomics"

PRINCIPAL AUTHOR

Name of Principal Author (Candidate)	Andres Mauricio Cifuentes Bernal
Contribution to the Paper Brief description of your work in this publication	Andres Cifuentes (A.C.). and Thuc Le (T.L.). conceived the experiments, A.C.conducted the experiments. A.C. and T.L analysed the results. A.C. and T.L. wrote the manuscript. Lin Liu (L.L.) and Jiuyong Li (J.L.) reviewed the manuscript.
Overall Percentage (%)	65%
Certification	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature *Signatures will be redacted by the Library at the time of publication*	Andres Cifuentes Digitally signed by Andres Cifuentes Date: 2023.11.22 13:59:45 +10'30'
Date	22-Nov-23

CO-AUTHOR CONTRIBUTIONS

By signing the Statement of Authorship, each author certifies that:

I. the candidate's stated contribution to the publication is accurate (as detailed above);

II. permission is granted for the candidate in include the publication in the thesis; and

III. the sum of all co-author contribution is equal to 100% less the candidate's stated contribution.

Name of Co-Author 1	Lin Liu

University of South Australia

% Contribution to the Paper	7.5%	
Signature *Signatures will be redacted by the Library at the time of publication*	Lin Liu Digitally signed by Lin Liu Date: 2023.11.22 20:53:41 +10'30'	22/11/2023 Date
Name of Co-Author 2	Jiuyong Li	
% Contribution to the Paper	7.5%	
Signature *Signatures will be redacted by the Library at the time of publication*		27/11/2023 Date
Name of Co-Author 3	Thuc Duy Le	
% Contribution to the Paper	20%	
Signature *Signatures will be redacted by the Library at the time of publication*		22/11/23 Date
Name of Co-Author 4		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 5		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 6		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 7		
% Contribution to the Paper		

University of South Australia

Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 8		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 9		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date
Name of Co-Author 10		
% Contribution to the Paper		
Signature *Signatures will be redacted by the Library at the time of publication*		Date

Bibliography

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] World Health Organization. Cancer Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/cancer, February 2022. Accessed: 2022-11-23.

[3] David Crosby, Sangeeta Bhatia, Kevin M Brindle, Lisa M Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C Fitzgerald, Sanjiv S Gambhir, Peter Kuhn, et al. Early detection of cancer. *Science*, 375(6586):eaay9040, 2022.

[4] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[5] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[6] Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[7] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. 152:1237–1251, 2013.

[8] Nastaran Zahir, Ruping Sun, Daniel Gallahan, Robert A Gatenby, and Christina Curtis. Characterizing the ecological and evolutionary dynamics of cancer. *Nature genetics*, 52(8):759–767, 2020.

[9] Franziska Michor, Yoh Iwasa, and Martin A Nowak. Dynamics of cancer progression. *Nature reviews cancer*, 4(3):197–205, 2004.

[10] Priyanka Kanth and John M Inadomi. Screening and prevention of colorectal cancer. *Bmj*, 374, 2021.

[11] Ahmed M Hassan and Magda El-Shenawee. Review of electromagnetic techniques for breast cancer detection. *IEEE Reviews in Biomedical Engineering*, 4:103–118, 2011.

[12] Véronique Bouvard, Nicolas Wentzensen, Anne Mackie, Johannes Berkhof, Julia Brotherton, Paolo Giorgi-Rossi, Rachel Kupets, Robert Smith, Silvina Arrossi, Karima Bendahhou, et al. The iarc perspective on cervical cancer screening. *New England Journal of Medicine*, 385(20):1908–1918, 2021.

[13] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[14] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385, 2018.

[15] Peter Priestley, Jonathan Baber, Martijn P Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, 575(7781):210–216, 2019.

[16] Aniruddha Chatterjee, Euan J Rodger, and Michael R Eccles. Epigenetic drivers of tumourigenesis and cancer metastasis. In *Seminars in cancer biology*, volume 51, pages 149–159. Elsevier, 2018.

[17] Jae Eun Lee and Mi-Young Kim. Cancer epigenetics: Past, present and future. In *Seminars in Cancer Biology*. Elsevier, 2021.

[18] G. S. Miglani. *Gene Expression*. Alpha Science International, New Delhi, INDIA, 2013.

[19] NHGRI. National Human Gennome Research Institute - Gene Expression, September 2022.

[20] Nature Education. Gene Expression | Learn Science at Scitable.

[21] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[22] John S Mattick and Igor V Makunin. Non-coding rna. *Human molecular genetics*, 15(suppl_1):R17–R29, 2006.

[23] Frank J Slack. Regulatory rnas and the demise of'junk'dna, 2006.

[24] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[25] Michael P Snyder, Thomas R Gingeras, Jill E Moore, Zhiping Weng, Mark B Gerstein, Bing Ren, Ross C Hardison, John A Stamatoyannopoulos, Brenton R Graveley, et al. Perspectives on encode. *Nature*, 583(7818):693–698, 2020.

[26] Frank J. Slack and Arul M. Chinnaiyan. The role of non-coding rnas in oncology. *Cell*, 179(5):1033–1055, 2019.

[27] Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, et al. A microrna expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences*, 103(7):2257–2261, 2006.

[28] Milena S Nicoloso, Riccardo Spizzo, Masayoshi Shimizu, Simona Rossi, and George A Calin. Micrornas—the micro steering wheel of tumour metastases. *Nature Reviews Cancer*, 9(4):293–302, 2009.

[29] Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, Julian M Hess, Randi Istrup Juul, Ziao Lin, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020.

[30] Yu Amanda Guo, Mei Mei Chang, and Anders Jacobsen Skanderup. Mutspot: detection of non-coding mutation hotspots in cancer genomes. *NPJ genomic medicine*, 5(1):1–5, 2020.

[31] Florian Kopp and Joshua T Mendell. Functional classification and experimental dissection of long noncoding rnas. *Cell*, 172(3):393–407, 2018.

[32] Xiaopei Zhang, Wei Wang, Weidong Zhu, Jie Dong, Yingying Cheng, Zujun Yin, and Fafu Shen. Mechanisms and functions of long non-coding rnas at multiple regulatory levels. *International journal of molecular sciences*, 20(22):5573, 2019.

[33] Tong Ihn Lee and Richard A Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.

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

[35] Pingyu Liu, Yijun Wang, and Xin Li. Targeting the untargetable kras in cancer therapy. *Acta Pharmaceutica Sinica B*, 9(5):871–879, 2019.

[36] Robert A Weinberg. Tumor suppressor genes. *Science*, 254(5035):1138–1146, 1991.

[37] Andrew O Giacomelli, Xiaoping Yang, Robert E Lintner, James M McFarland, Marc Duby, Jaegil Kim, Thomas P Howard, David Y Takeda, Seav Huong Ly, Eejung Kim, et al. Mutational processes shape the landscape of tp53 mutations in human cancer. *Nature genetics*, 50(10):1381–1387, 2018.

[38] Stephen P Jackson and Jiri Bartek. The dna-damage response in human biology and disease. *Nature*, 461(7267):1071–1078, 2009.

[39] Penny A Jeggo, Laurence H Pearl, and Antony M Carr. Dna repair, genome stability and cancer: a historical perspective. *Nature Reviews Cancer*, 16(1):35–42, 2016.

[40] Ashok R Venkitaraman. Cancer susceptibility and the functions of brca1 and brca2. *Cell*, 108(2):171–182, 2002.

[41] Carlo M Croce. Causes and consequences of microrna dysregulation in cancer. *Nature reviews genetics*, 10(10):704–714, 2009.

[42] Arianna Bottoni, Maria Chiara Zatelli, Manuela Ferracin, Federico Tagliati, Daniela Piccin, Cristina Vignali, George A Calin, Massimo Negrini, Carlo M Croce, and Ettore C Degli Uberti. Identification of differentially expressed micrornas by microarray: a possible role for microrna genes in pituitary adenomas. *Journal of cellular physiology*, 210(2):370–377, 2007.

[43] H Kang, C Kim, H Lee, JG Rho, JW Seo, Jin-Wu Nam, WK Song, SW Nam, W Kim, and EK Lee. Downregulation of microrna-362-3p and microrna-329 promotes tumor progression in human breast cancer. *Cell Death & Differentiation*, 23(3):484–495, 2016.

[44] Changchun Xiao and Klaus Rajewsky. Microrna control in the immune system: basic principles. *Cell*, 136(1):26–36, 2009.

[45] Samik Chakraborty, Minakshi Mazumdar, Shravanti Mukherjee, Pushpak Bhattacharjee, Arghya Adhikary, Argha Manna, Sreeparna Chakraborty, Poulami Khan, Aparna Sen, and Tanya Das. Restoration of p53/mir-34a regulatory axis decreases survival advantage and ensures bax-dependent apoptosis of non-small cell lung carcinoma cells. *FEBS letters*, 588(4):549–559, 2014.

[46] Melanie Ehrlich. Dna hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259, 2009.

[47] Xiaoya Duan, Yuanyong Huang, Xiaoxing Chen, Wencai Wang, Jiwei Chen, Jialun Li, Wei Yang, Jiwen Li, Qihan Wu, and Jiemin Wong. Moderate dna hypomethylation suppresses intestinal tumorigenesis by promoting caspase-3 expression and apoptosis. *Oncogenesis*, 10(5):38, 2021.

[48] Katsuhiko Ogata. *System Dynamics*. Pearson, fourth edition edition, 2004.

[49] Katsuhiko Ogata et al. *Modern control engineering*, volume 5. Prentice hall Upper Saddle River, NJ, 2010.

[50] Olga Korosteleva. *Stochastic Processes with R: An Introduction*. CRC Press, 2022.

[51] Robert G Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[52] Christian Gourieroux, Alain Monfort, and Gourieroux Christian. *Time series and dynamic models*, volume 3. Cambridge University Press, 1997.

[53] Elisabeth Roesch and Michael PH Stumpf. Parameter inference in dynamical systems with co-dimension 1 bifurcations. *Royal Society Open Science*, 6(10):190747, 2019.

[54] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[55] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[56] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[57] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

[58] Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. 2014.

[59] David A Lagnado and Steven Sloman. Learning causal structure. In *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society*, pages 560–565. Routledge, 2019.

[60] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, dec 2018.

[61] Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

[62] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424, aug 1969.

[63] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[64] David Stephen Geoffrey Pollock, Richard C Green, and Truong Nguyen. *Handbook of time series analysis, signal processing, and dynamics*. Elsevier, 1999.

[65] Alessio Moneta, Nadine Chlaß, Doris Entner, and Patrik Hoyer. Causal search in structural vector autoregressive models. In *NIPS Mini-Symposium on Causality in Time Series*, pages 95–114. PMLR, 2011.

[66] Lutz Kilian and Helmut Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

[67] Hans-Martin Krolzig. General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics*, 65:769–801, 2003.

[68] Peng Jiang, Sanju Sinha, Kenneth Aldape, Sridhar Hannenhalli, Cenk Sahinalp, and Eytan Ruppin. Big data in basic and translational cancer research. *Nature Reviews Cancer*, 22(11):625–639, 2022.

[69] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[70] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[71] Ghaneya Al-Khadairi, Adviti Naik, Remy Thomas, Boshra Al-Sulaiti, Shaheen Rizly, and Julie Decock. Prame promotes epithelial-to-mesenchymal transition in triple negative breast cancer. *Journal of translational medicine*, 17(1):1–14, 2019.

[72] Xiaowei Chen, Frank J. Slack, and Hongyu Zhao. Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions. *Bioinformatics*, 29(17):2137–2145, 06 2013.

[73] Fan Zheng, Marcus R Kelly, Dana J Ramms, Marissa L Heintschel, Kai Tao, Beril Tutuncuoglu, John J Lee, Keiichiro Ono, Helene Foussard, Michael

Chen, et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, 374(6563):eabf3067, 2021.

[74] Mandar S Chaudhary, Vu VH Pham, and Thuc D Le. Nibna: a network-based node importance approach for identifying breast cancer drivers. *Bioinformatics*, 37(17):2521–2528, 2021.

[75] Vu VH Pham, Lin Liu, Cameron P Bracken, Gregory J Goodall, Qi Long, Jiuyong Li, and Thuc D Le. Cbna: a control theory based method for identifying coding and non-coding cancer drivers. *PLoS computational biology*, 15(12):e1007538, 2019.

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

[77] Nayi Wang, Ji Zheng, Zhuo Chen, Yang Liu, Burak Dura, Minsuk Kwak, Juliana Xavier-Ferrucio, Yi-Chien Lu, Miaomiao Zhang, Christine Roden, et al. Single-cell microrna-mrna co-sequencing reveals non-genetic heterogeneity and mechanisms of microrna regulation. *Nature communications*, 10(1):1–12, 2019.

[78] Yuelei Zhang, Xiao Chang, and Xiaoping Liu. Inference of gene regulatory networks using pseudo-time series data. *Bioinformatics*, 37(16):2423–2431, 2021.

[79] Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

[80] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[81] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 11 2017.

[82] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[83] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[84] Kieran R. Campbell and Christopher Yau. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nature Communications*, 9(1), jun 2018.

[85] Xiaoqiang Sun, Ji Zhang, and Qing Nie. Inferring latent temporal progression and regulatory networks from cross-sectional transcriptomic data of cancer samples. *PLoS computational biology*, 17(3):e1008379, 2021.

[86] Andres M Cifuentes-Bernal, Vu Vh Pham, Xiaomei Li, Lin Liu, Jiuyong Li, and Thuc Duy Le. A pseudotemporal causality approach to identifying mirna–mrna interactions during biological processes. *Bioinformatics*, 37(6):807–814, 2021.

[87] Andres M Cifuentes-Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, Jiuyong Li, and Thuc Duy Le. Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories. *Briefings in Functional Genomics*, 21(6):455–465, 2022.

[88] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[89] John E Reid and Lorenz Wernisch. Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, 32(19):2973–2980, 2016.

[90] Kieran Campbell. *switchde: Switch-like differential expression across single-cell trajectories*, 2019. R package version 1.10.0.

[91] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics (Oxford, England)*, 31:3016–3026, September 2015.

[92] Rossin Erbe, Jessica Gore, Kelly Gemmill, Daria A Gaykalova, and Elana J Fertig. The use of machine learning to discover regulatory networks controlling biological systems. *Molecular cell*, 2022.

[93] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

[94] Hendrik Hache, Hans Lehrach, and Ralf Herwig. Reverse engineering of gene regulatory networks: a comparative study. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009:1–12, 2009.

[95] Fernando M Delgado and Francisco Gómez-Vela. Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial intelligence in medicine*, 95:133–145, 2019.

[96] Emily F Davis-Marcisak, Atul Deshpande, Genevieve L Stein-O'Brien, Won J Ho, Daniel Laheru, Elizabeth M Jaffee, Elana J Fertig, and Luciane T Kagohara. From bench to bedside: Single-cell analysis for cancer immunotherapy. *Cancer Cell*, 39(8):1062–1080, 2021.

[97] Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Biocomputing*, volume 3, 1998.

[98] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. BioMed Central, 2006.

[99] Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC bioinformatics*, 11(1):1–15, 2010.

[100] Francisco Gómez-Vela, Carlos D Barranco, and Norberto Díaz-Díaz. Incorporating biological knowledge for construction of fuzzy networks of gene associations. *Applied Soft Computing*, 42:144–155, 2016.

[101] Shohag Barman and Yung-Keun Kwon. A novel mutual information-based boolean network inference method from time-series gene expression data. *PloS one*, 12(2):e0171097, 2017.

[102] Hyejin Kang, Hongryul Ahn, Kyuri Jo, Minsik Oh, and Sun Kim. mirtime: identifying condition-specific targets of microrna in time-series transcript data using gaussian process model and spherical vector clustering. *Bioinformatics*, 37(11):1544–1553, 05 2021.

[103] Xin Zhou and Xiaodong Cai. Inference of differential gene regulatory networks based on gene expression and genetic perturbation data. *Bioinformatics*, 36(1):197–204, 2020.

[104] Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16:273, August 2015.

[105] Yuriy Sverchkov and Mark Craven. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS computational biology*, 13(6):e1005466, 2017.

[106] Shohag Barman and Yung-Keun Kwon. A boolean network inference from time-series gene expression data using a genetic algorithm. *Bioinformatics*, 34(17):i927–i933, 2018.

[107] Alicia T Specht and Jun Li. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. *Bioinformatics*, 33(5):764–766, 2017.

[108] Mengyuan Zhao, Wenying He, Jijun Tang, Quan Zou, and Fei Guo. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Briefings in Bioinformatics*, 22(5):bbab009, 2021.

[109] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, Yvan Saeys, and Pierre Geurts. Inferring gene regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, sep 2010.

[110] M Rubiolo, D H Milone, and G Stegmayer. Extreme learning machines for reverse engineering of gene regulatory networks from expression time series. *Bioinformatics*, 34(7):1253–1260, 11 2017.

[111] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 11 2018.

[112] Daniel Osorio, Yan Zhong, Guanxun Li, Qian Xu, Yongjian Yang, Yanan Tian, Robert S. Chapkin, Jianhua Z. Huang, and James J. Cai. sctenifoldknk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. *Patterns*, 3(3):100434, 2022.

[113] Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad, Safaai Deris, and Zalmiyah Zakaria. A review on the computational approaches for gene regulatory network construction. *Computers in biology and medicine*, 48:55–65, 2014.

[114] Yue Deng, Hector Zenil, Jesper Tegnér, and Narsis A Kiani. Hidi: an efficient reverse engineering schema for large-scale dynamic regulatory network reconstruction using adaptive differentiation. *Bioinformatics*, 33(24):3964–3972, 2017.

[115] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, 2011.

[116] Vân Anh Huynh-Thu and Pierre Geurts. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific reports*, 8(1):3384, 2018.

[117] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[118] Gary Hak Fui Tam, Chunqi Chang, and Yeung Sam Hung. Gene regulatory network discovery using pairwise granger causality. *IET systems biology*, 7(5):195–204, 2013.

[119] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[120] Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16(1):1–18, 2015.

[121] Daniel E Carlin, Evan O Paull, Kiley Graim, Christopher K Wong, Adrian Bivol, Peter Ryabinin, Kyle Ellrott, Artem Sokolov, and Joshua M Stuart. Prophetic granger causality to infer gene regulatory networks. *PloS one*, 12(12):e0170340, 2017.

[122] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

[123] Nan Papili Gao, SM Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, 2018.

[124] Xiaojie Qiu, Arman Rahimzamani, Li Wang, Bingcheng Ren, Qi Mao, Timothy Durham, José L McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell systems*, 10(3):265–274, 2020.

[125] Thuc D Le, Lin Liu, Bing Liu, Anna Tsykin, Gregory J Goodall, Kenji Satou, and Jiuyong Li. Inferring microrna and transcription factor regulatory networks in heterogeneous data. *BMC bioinformatics*, 14(1):1–13, 2013.

[126] Alexander Hartemink et al. Banjo: Bayesian network inference with java objects, 2005.

[127] Haijun Gong, Jakob Klinger, Kevin Damazyn, Xiangrui Li, and Shiyang Huang. A novel procedure for statistical inference and verification of gene regulatory subnetwork. *BMC bioinformatics*, 16(7):1–10, 2015.

[128] Xi Chen, Jinghua Gu, Xiao Wang, Jin-Gyoung Jung, Tian-Li Wang, Leena Hilakivi-Clarke, Robert Clarke, and Jianhua Xuan. Crnet: an efficient sampling approach to infer functional regulatory networks by integrating large-scale chip-seq and time-course rna-seq data. *Bioinformatics*, 34(10):1733–1740, 2018.

[129] Manuel Sanchez-Castillo, David Blanco, Isabel M Tienda-Luna, MC Carrion, and Yufei Huang. A bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics*, 34(6):964–970, 2018.

[130] Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li. Inferring microrna–mrna causal regulatory relationships from expression data. *Bioinformatics*, 29(6):765–771, 2013.

[131] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

[132] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[133] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, aug 2015.

[134] Nathan Wong and Xiaowei Wang. mirdb: an online resource for microrna target prediction and functional annotations. *Nucleic acids research*, 43(D1):D146–D152, 2014.

[135] Xiaowei Wang. Improving microrna target prediction by modeling with unambiguously identified microrna-target pairs from clip-ligation studies. *Bioinformatics*, 32(9):1316–1322, 2016.

[136] Cesaré Ovando-Vázquez, Daniel Lepe-Soltero, and Cei Abreu-Goodger. Improving microrna target prediction with gene expression profiles. *BMC genomics*, 17(1):364, 2016.

[137] Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, et al. mirtarbase update 2018: a resource for experimentally validated microrna-target interactions. *Nucleic acids research*, 46(D1):D296–D302, 2017.

[138] Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7):491–501, 2021.

[139] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.

[140] Vu Viet Hoang Pham, Lin Liu, Cameron Bracken, Gregory Goodall, Jiuyong Li, and Thuc Duy Le. Computational methods for cancer driver discovery: A survey. *Theranostics*, 11(11):5553, 2021.

[141] Li Ding, Gad Getz, David A Wheeler, Elaine R Mardis, Michael D McLellan, Kristian Cibulskis, Carrie Sougnez, Heidi Greulich, Donna M Muzny, Margaret B Morgan, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075, 2008.

[142] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, Matthew B Callaway, David Dooling, Elaine R Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

[143] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214, 2013.

[144] Yoo-Ah Kim, Sanna Madan, and Teresa M. Przytycka. Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, page btw242, may 2016.

[145] David Tamborero, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244, 2013.

[146] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S

Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505:495–501, January 2014.

[147] Peilin Jia, Quan Wang, Qingxia Chen, Katherine E Hutchinson, William Pao, and Zhongming Zhao. Msea: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome biology*, 15(10):1–16, 2014.

[148] Julian M Hess, Andre Bernards, Jaegil Kim, Mendy Miller, Amaro Taylor-Weiner, Nicholas J Haradhvala, Michael S Lawrence, and Gad Getz. Passenger hotspot mutations in cancer. *Cancer cell*, 36(3):288–301, 2019.

[149] Abel Gonzalez-Perez and Nuria Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic acids research*, 40(21):e169–e169, 2012.

[150] Yi Han, Juze Yang, Xinyi Qian, Wei-Chung Cheng, Shu-Hsuan Liu, Xing Hua, Liyuan Zhou, Yaning Yang, Qingbiao Wu, Pengyuan Liu, et al. Driverml: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic acids research*, 47(8):e45–e45, 2019.

[151] Malvika Sudhakar, Raghunathan Rengaswamy, and Karthik Raman. Novel ratio-metric features enable the identification of new driver genes across cancer types. *Scientific reports*, 12(1):1–12, 2022.

[152] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.

[153] Renan Andrades and Mariana Recamonde-Mendoza. Machine learning methods for prediction of cancer driver genes: a survey paper. *Briefings in Bioinformatics*, 23(3):bbac062, 2022.

[154] Kyuri Jo, Inuk Jung, Ji Hwan Moon, and Sun Kim. Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. *Bioinformatics*, 32(12):i128–i136, jun 2016.

[155] Arunachalam Vinayagam, Travis E Gibson, Ho-Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang-Yu Liu, Norbert Perrimon, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, 113(18):4976–4981, 2016.

[156] Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

[157] Gal Dinstag and Ron Shamir. Prodigy: personalized prioritization of driver genes. *Bioinformatics (Oxford, England)*, November 2019.

[158] Thanos P Mourikis, Lorena Benedetti, Elizabeth Foxall, Damjan Temelkovski, Joel Nulsen, Juliane Perner, Matteo Cereda, Jesper Lagergren, Michael Howell, Christopher Yau, et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nature Communications*, 10(1):3101, 2019.

[159] Joel Nulsen, Hrvoje Misetic, Christopher Yau, and Francesca D Ciccarelli. Pan-cancer detection of driver genes at the single-patient resolution. *Genome medicine*, 13:1–14, 2021.

[160] Loes Olde Loohuis, Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9:e108358, 2014.

[161] Patrick Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.

[162] Keegan D Korthauer and Christina Kendziorski. Madgic: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, 31(10):1526–1535, 2015.

[163] Chunhui Cai, Gregory F Cooper, Kevin N Lu, Xiaojun Ma, Shuping Xu, Zhenlong Zhao, Xueer Chen, Yifan Xue, Adrian V Lee, Nathan Clark, Vicky Chen, Songjian Lu, Lujia Chen, Liyue Yu, Harry S Hochheiser, Xia Jiang, Q Jane Wang, and Xinghua Lu. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLoS computational biology*, 15:e1007088, July 2019.

[164] Joseph Cursons, Katherine A. Pillman, Kaitlin G. Scheer, Philip A. Gregory, Momeneh Foroutan, Soroor Hediyeh-Zadeh, John Toubia, Edmund J. Crampin, Gregory J. Goodall, Cameron P. Bracken, and Melissa J. Davis. Combinatorial targeting by micrornas co-ordinates post-transcriptional control of emt. *Cell Systems*, 7(1):77 – 91.e7, 2018.

[165] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

[166] Junhua Zhang, Shihua Zhang, Yong Wang, and Xiang-Sun Zhang. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC systems biology*, 7:1–12, 2013.

[167] Yong Wang and Yu Xia. Condition specific subnetwork identification using an optimization model. *Optimization and Systems Biology*, pages 333–340, 2008.

[168] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47:106–114, February 2015.

[169] Rafsan Ahmed, Ilyes Baali, Cesim Erten, Evis Hoxha, and Hilal Kazan. Mexcowalk: mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics*, 36(3):872–879, 2020.

[170] Vu VH Pham, Lin Liu, Cameron P Bracken, Gregory J Goodall, Jiuyong Li, and Thuc D Le. Drivergroup: a novel method for identifying driver gene groups. *Bioinformatics*, 36(Supplement_2):i583–i591, 2020.

[171] Philip A Gregory, Andrew G Bert, Emily L Paterson, Simon C Barry, Anna Tsykin, Gelareh Farshid, Mathew A Vadas, Yeesim Khew-Goodall, and Gregory J Goodall. The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting zeb1 and sip1. *Nature cell biology*, 10(5):593–601, 2008.

[172] Thuc Duy Le, Lin Liu, Junpeng Zhang, Bing Liu, and Jiuyong Li. From mirna regulation to mirna–tf co-regulation: computational approaches and challenges. *Briefings in bioinformatics*, 16(3):475–496, 2015.

[173] Junpeng Zhang, Thuc Duy Le, Lin Liu, Bing Liu, Jianfeng He, Gregory J Goodall, and Jiuyong Li. Inferring condition-specific mirna activity from matched mirna and mrna expression data. *Bioinformatics*, 30(21):3070–3077, 2014.

[174] Junpeng Zhang, Lin Liu, Jiuyong Li, and Thuc Duy Le. Lncmirsrn: identification and analysis of long non-coding rna related mirna sponge regulatory network in human cancer. *Bioinformatics*, 34(24):4232–4240, 2018.

[175] Junpeng Zhang, Thuc Duy Le, Lin Liu, and Jiuyong Li. Inferring mirna sponge co-regulation of protein-protein interactions in human breast cancer. *BMC bioinformatics*, 18(1):243, 2017.

[176] Eunjee Lee, Koichi Ito, Yong Zhao, Eric E. Schadt, Hanna Y. Irie, and Jun Zhu. Inferred miRNA activity identifies miRNA-mediated regulatory networks underlying multiple cancers. *BIOINFORMATICS*, 32(1):96–105, JAN 1 2016.

[177] Xing-Ming Zhao, Ke-Qin Liu, Guanghui Zhu, Feng He, Beatrice Duval, Jean-Michel Richer, De-Shuang Huang, Chang-Jun Jiang, Jin-Kao Hao, and Luonan Chen. Identifying cancer-related microRNAs based on gene expression data. *BIOINFORMATICS*, 31(8):1226–1234, APR 15 2015.

[178] Stefano Volinia, Rosa Visone, Marco Galasso, Elda Rossi, and Carlo M. Croce. Identification of microRNA activity by Targets' Reverse EXpression. *BIOINFORMATICS*, 26(1):91–97, JAN 1 2010.

[179] Liangqun Lu, Sara McCurdy, Sijia Huang, Xun Zhu, Karolina Peplowska, Maarit Tiirikainen, William A Boisvert, and Lana X Garmire. Time series mirna-mrna integrated analysis reveals critical mirnas and targets in macrophage polarization. *Scientific reports*, 6(1):1–14, 2016.

[180] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845, 2016.

[181] John Aldrich. AUTONOMY. *Oxford Economic Papers*, 41(1):15–34, 01 1989.

[182] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

[183] A Philip Dawid, Vanessa Didelez, et al. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

[184] M. Puchinskaya. 35p - patterns of n-cadherin (n-cad) and vimentin (vim) expression as a sign of epithelial-mesenchymal transition (emt) in prostate cancer (pca). *Annals of Oncology*, 30:vii11, 2019. Abstract Book of Molecular Analysis for Personalised Therapy (MAP) London, UK, 7–9 November 2019.

[185] Weiguo Shu, Diana Douglas, Sangeeta Kumari, Luis Romero, Luis G. Rodriguez, Chaozhong Zou, and Robert Newman. Abstract 1884: Development of a novel vim-rfp reporter line for colorectal cancer emt study and drug discovery. *Cancer Research*, 79(13 Supplement):1884–1884, 2019.

[186] Ching-Yi Liu, Hsi-Hui Lin, Ming-Jer Tang, and Yang-Kao Wang. Vimentin contributes to epithelial-mesenchymal transition cancer cell mechanics by mediating cytoskeletal organization and focal adhesion maturation. *Oncotarget*, 6(25965826):15966–15983, June 2015.

[187] Huiyu Xu, Yong Han, Jiaying Lou, Hongxian Zhang, Yue Zhao, Balazs Gyorffy, and Rong Li. Pdgfra, hsd17b4 and hmgb2 are potential therapeutic targets in polycystic ovarian syndrome and breast cancer. *ONCOTARGET*, 8(41):69520–69526, SEP 19 2017.

[188] Chih-Hung Chou, Nai-Wen Chang, Sirjana Shrestha, Sheng-Da Hsu, Yu-Ling Lin, Wei-Hsiang Lee, Chi-Dung Yang, Hsiao-Chin Hong, Ting-Yen Wei, Siang-Jyun Tu, et al. mirtarbase 2016: updates to the experimentally validated mirna-target interactions database. *Nucleic acids research*, 44(D1):D239–D247, 2015.

[189] Ioannis S Vlachos, Maria D Paraskevopoulou, Dimitra Karagkouni, Georgios Georgakilas, Thanasis Vergoulis, Ilias Kanellos, Ioannis-Laertis Anastasopoulos, Sofia Maniou, Konstantina Karathanou, Despina Kalfakakou, et al. Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. *Nucleic acids research*, 43(D1):D153–D159, 2014.

[190] Harsh Dweep and Norbert Gretz. mirwalk2. 0: a comprehensive atlas of microrna-target interactions. *Nature methods*, 12(8):697, 2015.

[191] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.

[192] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[193] Monika Drobna, Bronislawa Szarzynska-Zawadzka, Patrycja Daca-Roszak, Maria Kosmalska, Roman Jaksik, Michal Witt, and Malgorzata Dawidowska. Identification of Endogenous Control miRNAs for RT-qPCR in T-Cell Acute

Lymphoblastic Leukemia. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES*, 19(10), OCT 2018.

[194] Yazeed A. Al-Sheikh, Hazem K. Ghneim, Khalid K. Alharbi, and Mourad A. M. Aboul-Soud. Screening for differentially-expressed microRNA biomarkers in Saudi colorectal cancer patients by small RNA deep sequencing. *INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE*, 44(6):2027–2036, DEC 2019.

[195] John C. Bladen, Jun Wang, Ajanthah Sangaralingam, Mariya Moosajee, Caroline Fitchett, Claude Chelala, Michele Beaconsfield, Edel A. O'Toole, Michael P. Philpott, and Daniel G. Ezra. MicroRNA and transcriptome analysis in periocular Sebaceous Gland Carcinoma. *SCIENTIFIC REPORTS*, 8, MAY 14 2018.

[196] Katharina Uhr, Wendy J. C. Prager-van der Smissen, Anouk A. J. Heine, Bahar Ozturk, Marijn T. M. van Jaarsveld, Antonius W. M. Boersma, Agnes Jager, Erik A. C. Wiemer, Marcel Smid, John A. Foekens, and John W. M. Martens. MicroRNAs as possible indicators of drug sensitivity in breast cancer cell lines. *PLOS ONE*, 14(5), MAY 7 2019.

[197] Lin Wang, Miao-Jun Zhu, Ai-Min Ren, Hong-Fei Wu, Wu-Mei Han, Ruo-Ying Tan, and Rui-Qin Tu. A Ten-MicroRNA Signature Identified from a Genome-Wide MicroRNA Expression Profiling in Human Epithelial Ovarian Cancer. *PLOS ONE*, 9(5), MAY 9 2014.

[198] Ya-Ling Hsu, Jen-Yu Hung, Yen-Lung Lee, Feng-Wei Chen, Kuo-Feng Chang, Wei-An Chang, Ying-Ming Tsai, Inn-Wen Chong, and Po-Lin Kuo. Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. *ONCOTARGET*, 8(62):104831–104854, DEC 1 2017.

[199] Yu-hui He, Yi-sen Deng, Pan-xin Peng, Ning Wang, Jian-feng Wang, Zhenshan Ding, Xing Chen, and Xiao-feng Zhou. A novel messenger RNA and long noncoding RNA signature associated with the progression of nonmus-

cle invasive bladder cancer. *JOURNAL OF CELLULAR BIOCHEMISTRY*, 120(5):8101–8109, MAY 2019.

[200] Weiyu Xu, Si Yu, Jianping Xiong, Junyu Long, Yongchang Zheng, and Xinting Sang. CeRNA regulatory network-based analysis to study the roles of noncoding RNAs in the pathogenesis of intrahepatic cholangiocellular carcinoma. *AGING-US*, 12(2):1047–1086, JAN 31 2020.

[201] Wen-Ji Shangguan, Hai-Tao Liu, Zu-Jun Que, Fang-Fang Qian, Ling-Shuang Liu, and Jian-Hui Tian. TOB1-AS1 suppresses non-small cell lung cancer cell migration and invasion through a ceRNA network. *EXPERIMENTAL AND THERAPEUTIC MEDICINE*, 18(6):4249–4258, DEC 2019.

[202] Fehmida Bibi, Muhammad I. Naseer, Sana Akhtar Alvi, Muhammad Yasir, Asif A. Jiman-Fatani, Ali Sawan, Adel M. Abuzenadah, Mohammed H. Al-Qahtani, and Esam I. Azhar. microRNA analysis of gastric cancer patients from Saudi Arabian population. *BMC GENOMICS*, 17(9), NOV 30 2015.

[203] Thuc Duy Le, Junpeng Zhang, Lin Liu, Huawen Liu, and Jiuyong Li. mirlab: An r based dry lab for exploring mirna-mrna regulatory relationships. *PLoS ONE*, 10(12):e0145386, 2015.

[204] Tuan Zea Tan, Qing Hao Miow, Yoshio Miki, Tetsuo Noda, Seiichi Mori, Ruby Yun-Ju Huang, and Jean Paul Thiery. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO molecular medicine*, 6(10):1279–1293, 2014.

[205] Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013.

[206] Thuc Duy Le, Junpeng Zhang, Lin Liu, Buu Minh Thanh Truong, Shu Hu, Taosheng Xu, and Jiuyong Li. Identifying microrna targets in epithelial-mesenchymal transition using joint-intervention causal inference. In *Pro-*

ceedings of the 8th International Conference on Computational Systems-Biology and Bioinformatics, CSBio '17, page 34–41, New York, NY, USA, 2017. Association for Computing Machinery.

[207] Aziz Khan and Anthony Mathelier. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, 18(1):287, 2017.

[208] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

[209] Felix Dietlein, Donate Weghorn, Amaro Taylor-Weiner, André Richters, Brendan Reardon, David Liu, Eric S Lander, Eliezer M Van Allen, and Shamil R Sunyaev. Identification of cancer driver genes based on nucleotide context. *Nature genetics*, 52(2):208–218, 2020.

[210] Mohammadreza Mohaghegh Neyshabouri, Seong-Hwan Jun, and Jens Lagergren. Inferring tumor progression in large datasets. *PLoS computational biology*, 16(10):e1008183, 2020.

[211] Cameron P Bracken, Hamish S Scott, and Gregory J Goodall. A network-biology perspective of microrna function and dysfunction in cancer. *Nature Reviews Genetics*, 17(12):719–732, 2016.

[212] Hong-Mei Zhang, Shuzhen Kuang, Xushen Xiong, Tianliuyun Gao, Chenglin Liu, and An-Yuan Guo. Transcription factor and microrna co-regulatory loops: important regulatory motifs in biological processes and diseases. *Briefings in bioinformatics*, 16(1):45–58, 2015.

[213] Heiko Horn, Michael S Lawrence, Candace R Chouinard, Yashaswi Shrestha, Jessica Xin Hu, Elizabeth Worstell, Emily Shea, Nina Ilic, Eejung Kim, Atanas Kamburov, et al. Netsig: network-based discovery from cancer genomes. *Nature methods*, 15(1):61–66, 2018.

[214] Jean Fan, Kamil Slowikowski, and Fan Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*, 52(9):1452–1465, 2020.

[215] Ting Jin, Nam D Nguyen, Flaminia Talos, and Daifeng Wang. Ecmarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics*, 37(8):1115–1124, 2021.

[216] Yoshiaki Nishiyama, Yutaka Saikawa, and Nobuaki Nishiyama. Interaction between the immune system and acute myeloid leukemia: A model incorporating promotion of regulatory t cell expansion by leukemic cells. *Biosystems*, 165:99–105, 2018.

[217] Hai Yang, Qiang Wei, Xue Zhong, Hushan Yang, and Bingshan Li. Cancer driver gene discovery through an integrative genomics approach in a non-parametric bayesian framework. *Bioinformatics*, 33(4):483–490, 2017.

[218] K Campbell. phenopath: Genomic trajectories with heterogeneous genetic and environmental backgrounds, 2021. *R package version*, 1(0).

[219] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8(1):1–12, 2017.

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

[221] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, mar 2015.

[222] Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E Assmus, Miguel A Andrade-Navarro, and Erich E Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Science signaling*, 4(189):rs8–rs8, 2011.

[223] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[224] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[225] Steven L Scott and Hal R Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23, 2014.

[226] Jinwen Qiu, S Rao Jammalamadaka, and Ning Ning. Multivariate bayesian structural time series model. *J. Mach. Learn. Res.*, 19(1):2744–2776, 2018.

[227] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017.

[228] Lauri A. Aaltonen, Federico Abascal, Adam Abeshouse, Hiroyuki Aburatani, David J. Adams, Nishant Agrawal, Keun Soo Ahn, Sung-Min Ahn, Hiroshi Aikata, Rehan Akbani, Kadir C. Akdemir, Hikmat Al-Ahmadie, Sultan T. Al-Sedairy, Fatima Al-Shahrour, Malik Alawi, Monique Albert, Kenneth Aldape, Ludmil B. Alexandrov, Adrian Ally, Kathryn Alsop, Eva G. Alvarez, Fernanda Amary, Samirkumar B. Amin, Brice Aminou, Ole Ammerpohl, Matthew J. Anderson, Yeng Ang, Davide Antonello, Pavana Anur, Samuel Aparicio, Elizabeth L. Appelbaum, Yasuhito Arai, Axel Aretz, Koji Arihiro, Shun-ichi Ariizumi, Joshua Armenia, Laurent Arnould, Sylvia Asa, Yassen Assenov, Gurnit Atwal, Sietse Aukema, J. Todd Auman, Miriam

R. R. Aure, Philip Awadalla, Marta Aymerich, Gary D. Bader, Adrian Baez-Ortega, Matthew H. Bailey, Peter J. Bailey, Miruna Balasundaram, Saianand Balu, Pratiti Bandopadhayay, Rosamonde E. Banks, Stefano Barbi, Andrew P. Barbour, Jonathan Barenboim, Jill Barnholtz-Sloan, Hugh Barr, Elisabet Barrera, John Bartlett, Javier Bartolome, Claudio Bassi, Oliver F. Bathe, Daniel Baumhoer, Prashant Bavi, Stephen B. Baylin, Wojciech Bazant, Duncan Beardsmore, Timothy A. Beck, Sam Behjati, Andreas Behren, Beifang Niu, Cindy Bell, Sergi Beltran, Christopher Benz, Andrew Berchuck, Anke K. Bergmann, Erik N. Bergstrom, Benjamin P. Berman, Daniel M. Berney, Stephan H. Bernhart, Rameen Beroukhim, Mario Berrios, Samantha Bersani, Johanna Bertl, Miguel Betancourt, Vinayak Bhandari, Shriram G. Bhosle, Andrew V. Biankin, Matthias Bieg, Darell Bigner, Hans Binder, Ewan Birney, Michael Birrer, Nidhan K. Biswas, Bodil Bjerkehagen, Tom Bodenheimer, Lori Boice, Giada Bonizzato, Johann S. De Bono, Arnoud Boot, Moiz S. Bootwalla, Ake Borg, Arndt Borkhardt, Keith A. Boroevich, Ivan Borozan, Christoph Borst, Marcus Bosenberg, Mattia Bosio, Jacqueline Boultwood, Guillaume Bourque, Paul C. Boutros, G. Steven Bova, David T. Bowen, Reanne Bowlby, David D. L. Bowtell, Sandrine Boyault, Rich Boyce, Jeffrey Boyd, Alvis Brazma, Paul Brennan, Daniel S. Brewer, Arie B. Brinkman, Robert G. Bristow, Russell R. Broaddus, Jane E. Brock, Malcolm Brock, Annegien Broeks, Angela N. Brooks, Denise Brooks, Benedikt Brors, Søren Brunak, Timothy J. C. Bruxner, Alicia L. Bruzos, Alex Buchanan, Ivo Buchhalter, Christiane Buchholz, Susan Bullman, Hazel Burke, Birgit Burkhardt, Kathleen H. Burns, John Busanovich, Carlos D. Bustamante, Adam P. Butler, Atul J. Butte, Niall J. Byrne, Anne-Lise Børresen-Dale, Samantha J. Caesar-Johnson, Andy Cafferkey, Declan Cahill, Claudia Calabrese, Carlos Caldas, Fabien Calvo, Niedzica Camacho, Peter J. Campbell, Elias Campo, Cinzia Cantù, Shaolong Cao, Thomas E. Carey, Joana Carlevaro-Fita, Rebecca Carlsen, Ivana Cataldo, Mario Cazzola, Jonathan Cebon, Robert Cerfolio, Dianne E. Chadwick, Dimple Chakravarty, Don Chalmers, Calvin Wing Yiu Chan, Kin Chan, Michelle Chan-Seng-Yue, Vishal S. Chandan, David K. Chang, Stephen J. Chanock, Lorraine A. Chantrill, Aurélien Chateigner, Nilanjan Chatterjee, Kazuaki Chayama, Hsiao-Wei Chen, Jieming Chen, Ken Chen, Yiwen Chen, Zhaohong Chen, Andrew D. Cherniack, Jeremy Chien, Yoke-Eng Chiew, Suet-Feung Chin, Juok Cho, Sunghoon Cho, Jung Kyoon Choi, Wan Choi, Christine Chomienne, Zechen Chong, Su Pin Choo, Angela

Chou, Angelika N. Christ, Elizabeth L. Christie, Eric Chuah, Carrie Cibulskis, Kristian Cibulskis, Sara Cingarlini, Peter Clapham, Alexander Claviez, Sean Cleary, Nicole Cloonan, Marek Cmero, Colin C. Collins, Ashton A. Connor, Susanna L. Cooke, Colin S. Cooper, Leslie Cope, Vincenzo Corbo, Matthew G. Cordes, Stephen M. Cordner, Isidro Cortés-Ciriano, Kyle Covington, Prue A. Cowin, Brian Craft, David Craft, Chad J. Creighton, Yupeng Cun, Erin Curley, Ioana Cutcutache, Karolina Czajka, Bogdan Czerniak, Rebecca A. Dagg, Ludmila Danilova, Maria Vittoria Davi, Natalie R. Davidson, Helen Davies, Ian J. Davis, Brandi N. Davis-Dusenbery, Kevin J. Dawson, Francisco M. De La Vega, Ricardo De Paoli-Iseppi, Timothy Defreitas, Angelo P. Dei Tos, Olivier Delaneau, John A. Demchok, Jonas Demeulemeester, German M. Demidov, Deniz Demircioğlu, Nening M. Dennis, Robert E. Denroche, Stefan C. Dentro, Nikita Desai, Vikram Deshpande, Amit G. Deshwar, Christine Desmedt, Jordi Deu-Pons, Noreen Dhalla, Neesha C. Dhani, Priyanka Dhingra, Rajiv Dhir, Anthony DiBiase, Klev Diamanti, Li Ding, Shuai Ding, Huy Q. Dinh, Luc Dirix, HarshaVardhan Doddapaneni, Nilgun Donmez, Michelle T. Dow, Ronny Drapkin, Oliver Drechsel, Ruben M. Drews, Serge Serge, Tim Dudderidge, Ana Dueso-Barroso, Andrew J. Dunford, Michael Dunn, Lewis Jonathan Dursi, Fraser R. Duthie, Ken Dutton-Regester, Jenna Eagles, Douglas F. Easton, Stuart Edmonds, Paul A. Edwards, Sandra E. Edwards, Rosalind A. Eeles, Anna Ehinger, Juergen Eils, Roland Eils, Adel El-Naggar, Matthew Eldridge, Kyle Ellrott, Serap Erkek, Georgia Escaramis, Shadrielle M. G. Espiritu, Xavier Estivill, Dariush Etemadmoghadam, Jorunn E. Eyfjord, Bishoy M. Faltas, Daiming Fan, Yu Fan, William C. Faquin, Claudiu Farcas, Matteo Fassan, Aquila Fatima, and The ICGC/T. C. G. A. Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.

[229] Afnan Abu-Thuraia, Marie-Anne Goyette, Jonathan Boulais, Carine Delliaux, Chloé Apcher, Céline Schott, Rony Chidiac, Halil Bagci, Marie-Pier Thibault, Dominique Davidson, et al. Axl confers cell migration and invasion by hijacking a peak1-regulated focal adhesion protein network. *Nature communications*, 11(1):3586, 2020.

[230] Todd A Aguilera, Marjan Rafat, Laura Castellini, Hussein Shehade, Mihalis S Kariolis, Angela Bik-Yu Hui, Henning Stehr, Rie Von Eyben, Dadi Jiang, Les-

ley G Ellies, et al. Reprogramming the immunological microenvironment through radiation and targeting axl. *Nature communications*, 7(1):13898, 2016.

[231] Alison M Nagle, Kevin M Levine, Nilgun Tasdemir, Julie A Scott, Kara Burlbaugh, Justin Kehm, Tiffany A Katz, David N Boone, Britta M Jacobsen, Jennifer M Atkinson, et al. Loss of e-cadherin enhances igf1–igf1r pathway activation and sensitizes breast cancers to anti-igf1r/insr inhibitors. *Clinical Cancer Research*, 24(20):5165–5177, 2018.

[232] Bing Chen, Xiao Zeng, Yu He, Xixi Wang, Ziwei Liang, Jingjing Liu, Peng Zhang, Hongxia Zhu, Ningzhi Xu, and Shufang Liang. Stc2 promotes the epithelial-mesenchymal transition of colorectal cancer cells through akt-erk signaling pathways. *Oncotarget*, 7(44):71400, 2016.

[233] Ayse Nihan Kilinc, Nami Sugiyama, Ravi Kiran Reddy Kalathur, Helena Antoniadis, Huseyin Birogul, Dana Ishay-Ronen, Jason T George, Herbert Levine, Mohit Kumar Jolly, and Gerhard Christofori. Histone deacetylases, mbd3/nurd, and tet2 hydroxylase are crucial regulators of epithelial–mesenchymal plasticity and tumor metastasis. *Oncogene*, 39(7):1498–1513, 2020.

[234] Tom L Stephen, Kyle K Payne, Ricardo A Chaurio, Michael J Allegrezza, Hengrui Zhu, Jairo Perez-Sanz, Alfredo Perales-Puchalt, Jenny M Nguyen, Ana E Vara-Ailor, Evgeniy B Eruslanov, et al. Satb1 expression governs epigenetic repression of pd-1 in tumor-reactive t cells. *Immunity*, 46(1):51–64, 2017.

[235] Cheng-Cheng Liu, Hua Wang, Wei-da Wang, Liang Wang, Wen-Jian Liu, Jing-Hua Wang, Qi-Rong Geng, and Yue Lu. Eno2 promotes cell proliferation, glycolysis, and glucocorticoid-resistance in acute lymphoblastic leukemia. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology*, 46:1525–1535, 2018.

[236] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and

Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943, 2016.

[237] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.

[238] Mohamed Mounir, Marta Lucchetta, Tiago C Silva, Catharina Olsen, Gianluca Bontempi, Xi Chen, Houtan Noushmehr, Antonio Colaprico, and Elena Papaleo. New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. *PLoS computational biology*, 15(3):e1006701, 2019.

[239] Li Fang, Yunjin Li, Lu Ma, Qiyue Xu, Fei Tan, and Geng Chen. Grndb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic acids research*, 49(D1):D97–D103, 2021.

[240] Eran R Andrechek. Her2/neu tumorigenesis and metastasis is regulated by e2f activator transcription factors. *Oncogene*, 34(2):217–225, 2015.

[241] Malte Sandsveden, Signe Borgquist, Ann H Rosendahl, and Jonas Manjer. Low thyroid hormone receptor alpha-2 (thrα-2) tumor expression is associated with unfavorable tumor characteristics and high breast cancer mortality. *Breast Cancer Research*, 23(1):1–12, 2021.

[242] Kenji Watanabe, Shigeru Yamamoto, Syuiti Sakaguti, Keishiro Isayama, Masaaki Oka, Hiroaki Nagano, and Yoichi Mizukami. A novel somatic mutation of sin3a detected in breast cancer by whole-exome sequencing enhances cell proliferation through erα expression. *Scientific reports*, 8(1):16000, 2018.

[243] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen,

Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[244] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in bioinformatics*, 22(3):bbaa190, 2021.

[245] Andres Cifuentes Bernal, Lin Liu, Jiuyong Li, and Thuc Le. Identifying cooperative genes causing cancer progression with dynamic causal inference. *bioRxiv*, 2023.

[246] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020.

[247] Jie Lyu, Jingyi Jessica Li, Jianzhong Su, Fanglue Peng, Yiling Elaine Chen, Xinzhou Ge, and Wei Li. Dorge: Discovery of oncogenes and tumor suppressor genes using genetic and epigenetic features. *Science Advances*, 6(46), nov 2020.

[248] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017.

[249] Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

[250] Thomas G Cotter. Apoptosis and cancer: the genesis of a research field. *Nature Reviews Cancer*, 9(7):501–507, 2009.

[251] Niklas Pfister, Stefan Bauer, and Jonas Peters. CausalKinetiX: Learning Stable Structures in Kinetic Systems, June 2019.

[252] Shizhe Chen, Ali Shojaie, and Daniela M Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112(520):1697–1707, 2017.

[253] Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1):1–16, 2016.

[254] Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016.

[255] Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel JB Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, et al. Gene set knowledge discovery with enrichr. *Current protocols*, 1(3):e90, 2021.

[256] Lisa Dressler, Michele Bortolomeazzi, Mohamed Reda Keddar, Hrvoje Misetic, Giulia Sartini, Amelia Acha-Sagredo, Lucia Montorsi, Neshika Wijewardhane, Dimitra Repana, Joel Nulsen, et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (ncg) resource. *Genome biology*, 23(1):1–22, 2022.

[257] Karolina A Zielińska and Vladimir L Katanaev. The signaling duo cxcl12 and cxcr4: Chemokine fuel for breast cancer tumorigenesis. *Cancers*, 12(10):3071, 2020.

[258] PP Pietkiewicz, A Lutkowska, M Lianeri, and PP Jagodzinski. Tamoxifen epigenetically modulates cxcl12 expression in mcf-7 breast cancer cells. *Biomedicine & pharmacotherapy*, 64(1):54–57, 2010.

[259] Haruko Hayasaka, Junichi Yoshida, Yasutaka Kuroda, Akihiro Nishiguchi, Michiya Matsusaki, Kei Kishimoto, Hitoshi Nishimura, Mari Okada, Yuki

Shimomura, Daichi Kobayashi, et al. Cxcl12 promotes ccr7 ligand–mediated breast cancer cell invasion and migration toward lymphatic vessels. *Cancer science*, 113(4):1338, 2022.

[260] Jairam Krishnamurthy, Jingqin Luo, Rama Suresh, Foluso Ademuyiwa, Caron Rigden, Timothy Rearden, Katherine Clifton, Katherine Weilbaecher, Ashley Frith, Anna Roshal, Pavan K. Tandra, Mathew Cherian, Tracy Summa, Brittney Haas, Shana Thomas, Leonel Hernandez-Aya, Mattias Bergqvist, Lindsey Peterson, and Cynthia X. Ma. A phase ii trial of an alternative schedule of palbociclib and embedded serum tk1 analysis. *NPJ BREAST CANCER*, 8(1), MAR 21 2022.

[261] Marzia Del Re, Ilaria Bertolini, Stefania Crucitta, Lorenzo Fontanelli, Eleonora Rofi, Claudia De Angelis, Lucrezia Diodati, Diletta Cavallero, Giulia Gianfilippo, Barbara Salvadori, Stefano Fogli, Alfredo Falcone, Cristian Scatena, Antonio Giuseppe Naccarato, Manuela Roncella, Matteo Ghilli, Riccardo Morganti, Andrea Fontana, and Romano Danesi. Overexpression of tk1 and cdk9 in plasma-derived exosomes is associated with clinical resistance to cdk4/6 inhibitors in metastatic breast cancer patients. *BREAST CANCER RESEARCH AND TREATMENT*, 178(1):57–62, NOV 2019.

[262] Marzia Del Re, Andrea Fontana, Stefania Crucitta, Ilaria Bertolini, Eleonora Rofi, Claudia De Angelis, Lucrezia Diodati, Diletta Cavallero, Barbara Salvadori, Alfredo Falcone, Riccardo Morganti, and Romano Danesi. Correlation of expression of tk1 in plasma-derived exosomes with clinical response to cdk4/6 inhibitors in breast cancer. *JOURNAL OF CLINICAL ONCOLOGY*, 36(15, S), MAY 20 2018.

[263] Luyu Zheng, Qun Liu, Ruijun Li, Shibin Chen, Jingyu Tan, Lina Li, Xichen Dong, Changzhi Huang, Tao Wen, and Jian Liu. Targeting mdk abrogates ifn-γ-elicited metastasis incancers of various origins. *Frontiers in Oncology*, 12, 2022.

[264] Xu Han, Minghui Li, Jin Xu, Jingyue Fu, Xinyang Wang, Jingyi Wang, Tiansong Xia, Shui Wang, and Ge Ma. mir-1275 targets mdk/akt signaling to

inhibit breast cancer chemoresistance by lessening the properties of cancer stem cells. *International Journal of Biological Sciences*, 19(1):89–103, 2023.

[265] Alexander H Chung, Tina M Leisner, Gabrielle Dardis, Marissa Bivins, and Leslie V Parise. Potent treatment combinations with cib1 depletion in triple-negative breast cancer. *Cancer Research*, 78(13_Supplement):3967–3967, 2018.

[266] Alexander H Chung, Tina M Leisner, Gabrielle J Dardis, Marissa M Bivins, Alana L Keller, and Leslie V Parise. Cib1 depletion with docetaxel or trail enhances triple-negative breast cancer cell death. *Cancer cell international*, 19(1):1–15, 2019.

[267] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018.

[268] Thomas Lin Pedersen. *tidygraph: A Tidy API for Graph Manipulation*, 2022. https://tidygraph.data-imaginist.com, https://github.com/thomasp85/tidygraph.

[269] Kaikai Wang, Huailei Liu, Jiaqi Liu, Xiaoxiong Wang, Lei Teng, Jun Zhang, Yi Liu, Yizheng Yao, Jun Wang, Yuan Qu, et al. Il1rn mediates the suppressive effect of methionine deprivation on glioma proliferation. *Cancer Letters*, 454:146–157, 2019.

[270] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[271] Mark Edmonds, Xiaojian Ma, Siyuan Qi, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Theory-based causal transfer: Integrating instance-level induc-

tion and abstract-level structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1283–1291, 2020.

[272] Vu VH Pham, Lin Liu, Cameron P Bracken, Thin Nguyen, Gregory J Goodall, Jiuyong Li, and Thuc D Le. pdriver: a novel method for unravelling personalized coding and mirna cancer drivers. *Bioinformatics*, 37(19):3285–3292, 2021.

[273] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.