Integrating single-cell and spatial transcriptomics in a computational multiscale model of Triple Negative Breast Cancer



Pedro Victori

St. Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

To Marta

I believe that when I die I shall rot, and nothing of my ego will survive. I am not young and I love life. But I should scorn to shiver with terror at the thought of annihilation. Happiness is nonetheless true happiness because it must come to an end, nor do thought and love lose their value because they are not everlasting. Many a man has borne himself proudly on the scaffold; surely the same pride should teach us to think truly about man's place in the world. Even if the open windows of science at first make us shiver after the cosy indoor warmth of traditional humanizing myths, in the end the fresh air brings vigour, and the great spaces have a splendour of their own.

— Bertrand Russell

Acknowledgements

I would like to extend my warmest thanks to the following people:

First of all, to all the patients that donated samples without whom all oncological research would be impossible.

To my supervisor Professor Francesca Buffa for giving me an opportunity when nobody else would, and for granting me the independence to pursue the answers to the questions I was most interested in.

To my other supervisor Professor Tim Maughan for accepting to help me at such a late stage and his suggestions on the medical side of my thesis.

To Dr Paul Savage, Dr Alexis Blanchet-Cohen, Professor Jiannis Ragoussis and the rest of authors of Savage *et al.*[1] for kindly sharing their single-cell RNA-seq data with me.

To Dr Karen Sayal for allowing me access to her spatial RNA-seq data.

To all the people on *StackOverflow* whose technical knowledge never failed to save me from computational dead-ends.

To Clara Pavillet for her help, our scientific discussions and for sharing with me her computational analysis of the Slide-seq data.

To Dr Jayathilake Pahala Gedara for all his help with the chemical part of the model and our illuminating discussions about modelling.

To Dr Duncan Gascoyne for inspiring me to become a scientist, all those years ago.

To my friends Esteban, María, Alejandro, Carlos, Rocío and my brother Álvaro for checking on me and for their emotional support during the lonely months of thesis writing, and for giving me things to look forward to afterwards.

To Hannah, for her kindness, her thoughtfulness, and our life-affirming walks and conversations.

To my mother Pilar, my role-model and to whom I owe everything, who was always convinced that this is what I was called to do, but allowed me to discover it by myself.

And finally, to Marta, for her never-failing companionship, for believing in me when not even I did, for her unwavering support during the worst of times, and for being the best friend anyone could wish for. I would be nowhere without you.

Abstract

This thesis presents a multiscale computational model of tumour growth, *microCJ*. *microCJ* models cells as independent agents, each with their own genetic identity and state. For each cell in the simulation, a Boolean gene regulatory network is iterated, the outputs of which govern the actions of the cell. Cells can proliferate, apoptose, enter growth arrest, or necrose due to lack of oxygen. Substance diffusion in the tumour microenvironment is modelled as gradients of concentration in space, and these substances can be consumed and/or produced by cells. Cells consume oxygen, growth factors (as well as producing them) and can also be treated with drugs. Cells compete with each other for nutrients and physical space in which to grow. Cells can present mutations and they can evolve, allowing the emergence of tumour heterogeneity and intraevolution. *microCJ* can be used to study the effects that different combination of mutations and tissue contexts can have on tumour growth, as well as its relationship to the tumour microenvironment.

Through the use of another tool I created, *Ingres*, the gene regulatory networks can be fitted using transcriptomics data, thus translating the biological data from single-cell RNA-seq into an executable model that can help to make sense of the data by providing a confined, reproducible platform for testing and generation of hypotheses.

These tools were used together to model triple-negative breast cancer, a disease with an unmet need, focusing on the context of EGFR-targeted treatment. To do this, several transcriptomic datasets (from the literature and collaborators) were employed, including a TNBC dataset sensitive to EGFR-targeted therapy by Savage *et al.* [1], to fit *microCJ* and study their characteristics and behaviour under certain perturbations, such as drug treatment.

Contents

List of Figures **xiii**

List of Tables **xvii**

List of Abbreviations **xix**

1 Introduction **1**

1.1 Triple-negative breast cancer 3

1.1.1 TNBC heterogeneity 5

1.1.2 Overview of current and novel therapeutic strategies 9

1.1.2.1 Chemotherapy 9

1.1.2.2 Immunotherapy 11

1.1.2.3 Targeted therapy and biomarkers 13

1.2 Mathematical modelling in oncology 29

1.2.1 Phenomenological and mechanistic modelling approaches . . 32

1.2.2 Boolean gene regulatory networks 35

1.2.2.1 Synchronous networks 36

1.2.2.2 Asynchronous networks 37

1.2.2.3 Probabilistic Boolean networks 37

1.2.3 Modelling therapies against cancer 39

1.2.3.1 Classification of mathematical models 40

1.2.3.2 Ordinary differential equation models 40

1.2.3.3 Partial differential equation models 43

1.2.3.4 Agent-based models 45

1.2.3.5 Other examples of studies using mathematical modelling 48

1.2.4 Opening the cell's "black box" 49

1.2.5 Data explosion and open access repositories 50

1.3 Aims of this thesis 52

2 Analysis of single-cell and spatial transcriptomics 53
2.1 An EGFR-dependent tumour-initiating program in TNBC 53
2.1.1 Analysis 55
2.2 Spatial transcriptomics 70
2.3 Breast cancer single-cell RNA-seq atlas 76
2.4 Conclusions 83
2.5 Methods 84
2.5.1 Quantification and analysis of single cell RNA-seq from Savage *et al.* [1] 84
2.5.2 Quantification and analysis of single cell RNA-seq breast cancer atlas from Gambardella *et al.* [297] 85
2.5.3 Spatial RNA-seq data 85
2.5.3.1 Collection and preparation of tissue samples 85
2.5.3.2 Library preparation and sequencing 86
2.5.4 Quantification and analysis of spatial RNA-seq breast cancer samples 87

3 Network modelling 89
3.0.1 Attractors 90
3.1 The MAPK network 91
3.1.1 Results 94
3.2 The *Ingres* pipeline 105
3.2.1 Implementation 106
3.2.1.1 Infer node activation from RNA sequencing data . 106
3.2.1.2 Computing the Probabilistic Boolean Network . . . 107
3.2.1.3 Analysis 110
3.2.2 Demonstration 110
3.3 Conclusions 118
3.4 Methods 119
3.4.1 Network analyses 119
3.4.2 Ingres 120

4 Description of the computational tumour model 121
4.1 Features 124
4.1.1 Network modelling 124
4.1.1.1 Fate decision 126
4.1.1.2 Import transcriptional data 127
4.1.1.3 Random mutations 129
4.1.2 Cell placement 130
4.1.3 Diffusion 131

4.2 Results . . . 134
4.2.1 Effect of mutations on the model behaviour . . . 134
4.2.2 Competitive advantage in multiclonal simulations . . . 137
4.2.3 The role of the microenvironment in the model evolution . . 141
4.2.4 Probabilistic Boolean Networks for a more granular simulation 150
4.2.5 Acquired mutations and clonal evolution . . . 152
4.2.6 Drug treatment . . . 162
4.2.7 Sensitivity analyses . . . 168
4.2.7.1 Instant growth arrest . . . 168
4.2.7.2 Decision window . . . 168
4.2.7.3 Growth arrest duration . . . 172
4.2.7.4 Shoving . . . 173
4.2.7.5 Cell placement method . . . 174
4.3 Conclusions . . . 183
4.4 Methods . . . 185
4.4.1 Implementation . . . 185
4.4.2 CLI and GUI . . . 188
4.4.3 Settings . . . 189

5 Analysis of EGFR heterogeneity using *microCJ* with experimental and transcriptomics data 195
5.1 *In vitro* response . . . 196
5.2 An extended RTK module in the MAPK network . . . 203
5.3 *microCJ* simulations with new network . . . 209
5.4 Fitting *microCJ* with single-cell RNA-seq data . . . 215
5.5 Conclusions . . . 233
5.6 Methods . . . 240
5.6.1 *In vitro* methods . . . 240
5.6.1.1 Western blot . . . 240
5.6.1.2 Flow cytometry . . . 241
5.6.1.3 Proliferation assays . . . 241
5.6.2 Importing of transcriptomics data . . . 241

6 Discussion 243

Appendices

A Documentation and tutorial vignette for the *Ingres* package 249

B Settings for all *microCJ* figures in Chapter 4 281
B.1 Default settings . . . 281
B.2 Specific settings . . . 284

References **285**

List of Figures

1.1 Key signal transduction mechanisms in TNBC tumorigenesis 16
1.2 Biomarkers and cancer histologies associated with *BRCAness* and its potential clinical utility 17
1.3 EGFR signal transduction 23
1.4 A workflow for modelling biological systems 32
1.5 Phenomenological and mechanistic approaches to biological modelling, with their strengths and shortcomings 33

2.1 A graphic abstract of the study carried out by Savage *et al.* [1] 54
2.2 Quality control measurements 55
2.3 Overview of the datasets 56
2.4 Clustering tree of Savage *et al.* [1] data 59
2.5 Cell abundance per cluster in Savage *et al.* [1] data 60
2.6 Principal component analysis 61
2.7 Cell abundance per cell cycle phase in Savage *et al.* [1] data 62
2.8 EGFR status 63
2.9 Marker genes 65
2.10 Cancer pathways enrichment 67
2.11 Transcription factor activity score 67
2.12 Heatmaps showing the enrichment per cluster for two different signatures 68
2.13 TNBC subtype scoring 69
2.14 Quality control measurements 71
2.15 Clustering of Slide-seq data 73
2.16 TNBC subtype scoring 74
2.17 UMAP of annotated clusters 75
2.18 Quality control measurements 77
2.19 Overview of the data used from Gambardella *et al.* [297] 78
2.20 EGFR expression 79
2.21 Pathway enrichment 80
2.22 TNBC subtype signature for Gambardella *et al.* [297] dataset. 81
2.23 Savage *et al.* [1] cell type signature for the Gambardella *et al.* [297] dataset. 82

3.1 Types of attractors in synchronous and asynchronous Boolean networks 91
3.2 The MAPK Boolean network by Grieco *et al.*[313] 92
3.3 The modified MAPK network with added hypoxia module 93
3.4 Four-state attractors for the unperturbed/inputs-off synchronous MAPK network 94
3.5 Relationship between fate attractors and initial state node activation in the unperturbed-inputs-off synchronous network 96
3.6 The steady-state attractors for over-expressed EGFR 98
3.7 Synchronous 8-state attractors for over-expressed EGFR 99
3.8 Fates and initial states for over-expressed EGFR synchronous network 100
3.9 The steady-state attractors for over-expressed EGFR networks with additional perturbations 101
3.10 ERK disrupts the anti-proliferative effects of TGFβ 102
3.11 RAS-mediated proliferation only occurs in the absence of JNK . . . 104
3.12 A network generated by *Ingres* 109
3.13 Interactive app to explore the results 110
3.14 The network of the components involved in entry into apoptosis and cell cycle, through E2F1, by Remy *et al.*[336] 111
3.15 State transition computation for the networks produced by *Ingres* . 112
3.16 State transition computation for the networks produced by *Ingres* . 113
3.17 State transition computation for the networks produced by *Ingres* using a bladder cancer dataset 114
3.18 Networks produced by *Ingres* for each of the clusters using the bladder cancer dataset 115

4.1 Overview of *microCJ* functionality 123
4.2 *microCJ* can import spatial RNA-seq data 128
4.3 Populations with different mutations growing independently 135
4.4 Average node activation of independent populations across different time points 136
4.5 Effect of mutations on competitive performance 137
4.6 Average node activation of competing subpopulations across different time points 138
4.7 Growth of a heterogeneous tumour at several time points 139
4.8 Fate activation during the competition simulation 140
4.9 Growth comparison between normoxic and hypoxic monoclonal tumours 141
4.10 Growth comparison between normoxic and hypoxic multiclonal tumours 142
4.11 Oxygen diffusion through a 2D space hypoxic simulation 143
4.12 Average node activation of independent populations across different time points under hypoxia 144

4.13 Growth of a heterogeneous tumour under hypoxic conditions at several time points . . . 145
4.14 State of each cell in a heterogeneous tumour under hypoxic conditions at several time points . . . 146
4.15 EGF diffusion . . . 147
4.16 EGF diffusion through a 2D space simulation . . . 148
4.17 Effect of EGF production in hypoxic conditions . . . 149
4.18 Competition simulations using probabilistic networks . . . 151
4.19 The acquired mutations during a simulation with sustained DNA damage, with a control simulation . . . 156
4.20 Comparison between mutated genes in control and sustained DNA damage simulation . . . 157
4.21 Correlation between mutated genes in the DNA damage simulation and cell generation . . . 158
4.22 Mutations acquired by cell during a simulation with ongoing DNA damage . . . 161
4.23 Cells in the DNA damage simulation, coloured by generation . . . 161
4.24 Diffusion and intake of a timed drug treatment . . . 163
4.25 Effect of a EGFR-blocking drug on cell proliferation . . . 164
4.26 Effect of a EGFR-blocking drug on competition . . . 165
4.27 Decision window sensitivity analysis . . . 168
4.28 Growth arrest duration sensitivity analysis . . . 169
4.29 Instant growth arrest comparison . . . 169
4.30 Instant growth arrest relationship with overall fate activation . . . 170
4.31 Effects of the "shoving" parameter on the growth of a tumour . . . 171
4.32 3D random placement . . . 175
4.33 3D square placement . . . 176
4.34 2D hexagonal placement . . . 178
4.35 2D random placement . . . 180
4.36 2D square placement . . . 182
4.37 Overview of *microCJ* codebase structure . . . 186
4.38 A view of *microCJ* GUI . . . 187
4.39 Performance of microCJ . . . 188

5.1 Flow-cytometry of MDA-MB-468 cells labelled with an anti-EGFR primary antibody . . . 197
5.2 Expression of EGFR and p-EGFR in MDA-MB-231 and 468 cells treated with gefitinib . . . 198
5.3 Proliferation of MDA-MB-468 and MDA-MB-231 cells treated with gefitinib . . . 201

5.4 New nodes added to the MAPK network 204
5.5 8-state synchronous attractors for the unperturbed extended network 205
5.6 Attractors for the EGFR over-expressed extended network 206
5.7 8-state attractors for the p53- extended network 207
5.8 Attractors for positive drug inputs in the extended network 208
5.9 Attractors for different perturbations of the extended network . . . 208
5.10 Competition simulation with the new network 209
5.11 *microCJ* running the new extended network with different perturbations 210
5.12 Acquired mutations in the extended network under different treatments 212
5.13 *microCJ* mutation panel simulations 213
5.14 An overview of the PBNs produced by *Ingres*, one per each cell, with the dataset from Savage *et al.* [1] 218
5.15 Fitting the model with the control scRNA-seq dataset from Savage *et al.* [1] 220
5.16 Behaviour of the model when fitted with all datasets scRNA-seq data from Savage *et al.* [1] 221
5.17 Drug panel simulation with Savage *et al.* [1] data 222
5.18 Expression of top targets in the Savage *et al.* [1] scRNA-seq dataset . 224
5.19 An overview of the PBNs produced by *Ingres*, one per each cell, with the dataset from Gambardella *et al.* [297] 225
5.20 Behaviour of the model when fitted with the scRNA-seq data from Gambardella *et al.* [297] 225
5.21 Drug panel simulation with Gambardella *et al.* [297] data 226
5.22 Expression of top targets in the Gambardella *et al.* [297] scRNA-seq dataset 228
5.23 An overview of the PBNs produced by *Ingres*, one per each cell, with the Slide-seq spatial RNA-seq dataset (puck 3) 229
5.24 Behaviour of the model when fitted with the Slide-seq spatial RNA-seq dataset (puck three) 229
5.25 Drug panel simulation with the Slide-seq spatial RNA-seq dataset (puck 3) 230
5.26 Expression of top targets in the Slide-seq spatial RNA-seq dataset . 232

List of Tables

1.1 Genomic TNBC subtypes and potential therapeutic targets, adapted from Collignon *et al.*[5], data from Lehmann *et al.*[21] 6
1.2 Potential therapeutic strategies and targeted drugs used in particular subtypes of triple negative breast cancer, adapted from Yin *et al.*[8] . 10
1.3 Summary of prognostic biomarkers in TNBC, adapted from da Silva *et al.*[10] 14

2.1 Slide-seq Custom Primers 87

4.1 A full description of every setting in *microCJ* 193

5.1 Characteristics of triple-negative breast cancer cell lines used in the preliminary experiments 196
5.2 List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Savage *et al.*[1] dataset. 216
5.3 List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Gambardella *et al.*[297] dataset. 217
5.4 List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Slide-seq dataset. 230

B.1 General settings 282
B.2 Diffusion settings 283
B.3 Settings by figure 284
B.4 Diffusion settings for drugs (figure 4.25 and related) 284

List of Abbreviations

ABM . . . . . Agent based model.

BL1 . . . . . Basal-like 1.

BL2 . . . . . Basal-like 2.

BLIA . . . . . Basal-like immune-activated.

BLIS . . . . . Basal-like immunosuppressed.

CLI . . . . . Command line interface.

CT . . . . . Chemotherapy.

DFS . . . . . Disease-free survival.

DSS . . . . . Disease-specific survival.

EGFR . . . . . Epidermal growth factor receptor.

EMT . . . . . Epithelial-mesenchymal transition.

ER . . . . . Estrogen receptor.

GRN . . . . . Gene regulatory network.

GUI . . . . . Graphical user interface.

HER2 . . . . . Human epidermal growth factor receptor-2.

HRD . . . . . Homologous recombination deficiency.

HRR . . . . . Homologous recombination repair.

Ig . . . . . Inmunoglobulin.

IHC . . . . . Immunohistochemistry.

IM . . . . . Immunomodulatory.

LAR . . . . . Luminal androgen receptor.

lncRNA . . . . Long noncoding RNA .

M . . . . . Mesenchymal.

mAb . . . . . Monoclonal antibody.

miRNA . . . . microRNA.

mRNA . . . . Messenger RNA.

MSL . . . . . Mesenchymal stem-like.

NES . . . . . Normalised enrichment score.

NP . . . . . Nanoparticle.

NSCLC . . . . Non-small cell lung cancer.

ORR . . . . . Overall response rate.

OS . . . . . Overall survival.

PARP . . . . . Poly (ADP-ribose) polymerase.

PBN . . . . . Probabilistic Boolean network.

pCR . . . . . Pathologic complete response.

PD-1 . . . . . Programmed cell death protein 1.

PFS . . . . . Progression free survival.

PR . . . . . Progesterone receptor.

PRIMA-1 . . . p53 reactivation and induction of massive apoptosis 1.

RTK . . . . . Receptor tyrosine kinase.

TCGA . . . . . The cancer genome atlas.

TGF . . . . . Transforming growth factor.

TIC . . . . . Tumour-initiating cells.

TILs . . . . . Tumour-infiltrating lymphocites.

TKI . . . . . tyrosine kinase inhibitor.

TNBC . . . . . Triple-negative breast cancer.

UMI . . . . . unique molecular identifier.

WT . . . . . Wild type.

Published work

This thesis contains some work that has already been published. In every case, I was the first author and wrote the first draft. I have obtained permission from my co-authors to include parts of those works in this thesis. A list of all works included follows.

- Victori, P. & Buffa, F. M. The Many Faces of Mathematical Modelling in Oncology. *The British Journal of Radiology,* 20180856. doi:10.1259/bjr.20180856 (Nov. 2018)
- Victori, P. & Buffa, F. M. *Ingres: From Single-Cell RNA-seq Data to Single-Cell Probabilistic Boolean Networks* Sept. 2022. doi:10.1101/2022.09.04.506528 (Preprint)

But the story of leukemia —the story of cancer— isn't the story of doctors who struggle and survive, moving from institution to another. It is the story of patients who struggle and survive, moving from one embankment of illness to another. Resilience, inventiveness, and survivorship–qualities often ascribed to great physicians–are reflected qualities, emanating first from those who struggle with illness and only then mirrored by those who treat them. If the history of medicine is told through the stories of doctors, it is because their contributions stand in place of the more substantive heroism of their patients.

— Siddhartha Mukherjee, The Emperor of All Maladies: A Biography of Cancer

1 Introduction

Contents

1.1 Triple-negative breast cancer 3
1.1.1 TNBC heterogeneity 5
1.1.2 Overview of current and novel therapeutic strategies . . 9
1.1.2.1 Chemotherapy 9
1.1.2.1.1 Platinum-based chemotherapy 9
1.1.2.2 Immunotherapy 11
1.1.2.3 Targeted therapy and biomarkers 13
1.1.2.3.1 *TP53* 16
1.1.2.3.2 *BRCA1/2* mutation, "BRCAness" and PARP inhibitors 17
1.1.2.3.3 Ki-67 19
1.1.2.3.4 Androgen receptor 20
1.1.2.3.5 c-KIT (CD117) 20
1.1.2.3.6 microRNAs and long non-coding RNAs 21
1.1.2.3.7 PI3K/AKT/mTOR pathway 21
1.1.2.3.8 Vascular endothelial growth factor (VEGF) 22
1.1.2.3.9 Epidermal growth factor receptor (EGFR) 22
1.2 Mathematical modelling in oncology 29
1.2.1 Phenomenological and mechanistic modelling approaches 32
1.2.2 Boolean gene regulatory networks 35
1.2.2.1 Synchronous networks 36
1.2.2.2 Asynchronous networks 37
1.2.2.3 Probabilistic Boolean networks 37
1.2.3 Modelling therapies against cancer 39
1.2.3.1 Classification of mathematical models 40
1.2.3.2 Ordinary differential equation models 40
1.2.3.3 Partial differential equation models 43

1.2.3.4 Agent-based models 45
1.2.3.5 Other examples of studies using mathematical modelling 48
1.2.4 Opening the cell's "black box" 49
1.2.5 Data explosion and open access repositories 50
1.3 Aims of this thesis 52

The work presented in this thesis is made up of several independent pieces that are brought together at the end of the thesis. In chapter 2, I discuss the analysis of several triple negative breast cancer (TNBC) single-cell transcriptomics datasets. The main one among these comes from a study on a exceptional responder to the epidermal growth factor receptor (EGFR) targeted therapy[1], and because of this EGFR is an ongoing focus of this thesis. Chapter 3 is focused on network models. In it, I present a tool I have developed, *Ingres*, that takes single-cell RNA-seq datasets, such as the ones introduced in chapter 2, and fits Boolean gene regulatory network models with them, producing probabilistic Boolean networks that are tailored to every individual cell. Chapter 3 also discuss one of these Boolean networks, a model of the MAP kinases pathway, and presents several analysis performed on it. Chapter 4 presents the main result of my thesis, the modelling framework *microCJ*. *microCJ* is a multi-level model of cancer, and includes Boolean networks, such as the ones discussed in chapter 3, to simulate the intracellular signalling processes of each cell. It also models cell-cell interactions and competition, and the diffusible substances that are part of the tumour microenvironment. Finally, chapter 5 brings together all of this pieces: it fits the network introduced in chapter 3 with the transcriptomics data presented in chapter 2, using the *Ingres* pipeline, and describes how a *microCJ* model built upon these fitted networks behaves and responds to *in-silico* targeted therapy.

This introduction provides an overview of all the concepts needed to follow the rest of the thesis. It discuss TNBC and the broader context of breast cancer, TNBC subtypes, the current therapies and the known therapeutic biomarkers. It also summarises the history of mathematical modelling in biology, and goes over the most important modelling paradigms in computational oncology.

1.1 Triple-negative breast cancer

Breast cancer (BC) is the most frequently diagnosed cancer, accounting for almost one in four cancer cases among women worldwide[4]. This makes it the leading cause of cancer death among women. According to GLOBOCAN, the estimated worldwide incidence of breast cancer in 2018 was of 2,088,849 cases[4].

These statistics include all subtypes of breast cancer, but it is well known that this is not a homogeneous disease. Four major intrinsic subtypes have been identified by genomic studies: the luminal subtypes A and B, which express oestrogen and progesterone receptor (ER, PR)-related genes, basal-like (BL) BC, and HER2-positive BC[5].

Around 70% of all human breast tumours express ER and/or PR. These hormone receptors (HR) are the primary transcription factor driving oncogenesis in HR-positive cancer and are the target of hormonal therapy[6]. Their expression levels are a predictor of response to ER-targeted therapy and are used to classify breast cancer in the different subtypes mentioned above.

TNBC is a heterogeneous sub-type of breast cancer characterised by the absence of expression of oestrogen and progesterone receptors and human epidermal growth factor receptor-2 (HER2). The definition of negative oestrogen receptor status by immunohistochemistry (IHC) is inconsistent in the literature, with some definitions considering ER expression to be significant only if at least 10% of tumour cells express the receptor, while other institutions have defined TNBC as BC presenting fewer than 1% of tumour cells expressing ER and progesterone receptor via IHC[5].

TNBC patients have a shorter survival time compared to other breast cancer subtypes, with a mortality of 40% within five years after diagnosis. TNBC is highly invasive and has an aggressive clinical behaviour, with a higher risk of both local and distant relapses that frequently present as visceral and/or brain metastases[7]. In total, approximately 46% of TNBC patients will have distant metastasis, often as soon as three years after diagnosis. The average time to relapse is 19-40 months, compared to 35-67 months in non-TNBC patients. The median survival time after metastasis is only 13.3 months, and the recurrence rate after surgery is as high

as 25%. The mortality rate of TNBC patients within 3 months after recurrence is as high as 75%[8]. Late tumour recurrences are unusual in this subtype and recurrences are generally not observed after eight years[9]. At diagnosis, patients usually present clinically positive axillary lymph node and larger primary tumour size compared to other breast cancer subtypes[9,10].

TNBC accounts for 15-20% of all breast cancers. It mostly occurs in premenopausal women under 40 years old, and in women of African or Hispanic ancestry: a study of 57,483 breast cancer patients showed that African-American women and Hispanic women are twice as likely and 30% more likely, respectively, than white women to be diagnosed with TNBC[11]. Non-white women are also over-represented among TNBC patients and diagnosed at a younger age[12,13]. The association between TNBC molecular subtype (see section 1.1.1) and ethnicity is poorly researched. Despite the mentioned imbalance in patient ancestry, few TNBC classification studies[14] include non-white women: in the Cancer Genome Atlas (TCGA) breast cancer set of more than 1000 women, there are few samples from non-white patients — among those with TNBC, there are eight samples from Asian women, four from Hispanic ones, and 33 from African-Americans[14].

75-80% of TNBCs display a BL molecular phenotype. But not all BL breast cancers are TNBC: up to 54% of cancers in the BL subgroup do not have a TN phenotype[5] — some BL cancers express ER or overexpress HER2. Currently there is no international definition of TNBC/BL cancer, but the most appropriate approach to use in the absence of access to molecular profiling is a panel of four antibodies (ER, HER2, CK5/6, and EGFR HER1), which can best characterize BL tumours based on IHC criteria[5].

TNBC lacks a standard of care approach guided by tumour biology[15]. Due to its molecular phenotype, TNBC is not sensitive to endocrine therapy or most molecular targeted therapies. Therefore, chemotherapy (CT) is the main systemic treatment, but the efficacy of conventional postoperative adjuvant chemoradiotherapy is poor. Because of this, it is crucial to develop new targets for TNBC therapy.

TNBC is highly heterogeneous regarding treatment responses and long-term outcomes[16,17]. Some patients respond very well to neoadjuvant chemotherapy and present a pathological complete response* (pCR) at the time of surgery, while others present no response to chemotherapy and suffer from early relapse after sugery[5]. Unfortunately, predictive factors have not yet been identified for which patients will present a pCR or will benefit from CT at the time of diagnosis.

Since homologous recombination deficiency (HRD) is common amonst TNBC patients[18,19], therapies that exploit this defect, such as platinum compounds and PARP inhibitors, are a promising strategy for treatment. This will be further discussed in one of the following sections.

1.1.1 TNBC heterogeneity

The vast majority of TNBCs are high-grade invasive ductal carcinomas, but some rare cases are histologically different, such as adenoid cystic carcinoma, secretory carcinoma, medullary carcinoma, and metaplastic carcinoma. The prognosis depends on the TNBC pathological subtype. The most comprehensive and used subclassification was proposed in 2011 by Lehmann *et al.*[21], who divided TNBC into six subtypes through gene expression profiling of tumour samples from 587 TNBC patients. The six subtypes are basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal (M), mesenchymal stem-like (MSL), immunomodulatory (IM), and luminal androgen receptor (LAR). The strengths of this study were to further identify the molecular drivers in cell-line models to provide preclinical platforms for the development of effective therapies (table 1.1).

The BL1 subtype shows abnormal expression of cell cycle-regulating and DNA repair-related genes (high amplification of *MYC*, *PIK3CA*, *CDK6*, *AKT2*, *KRAS*, *FGFR1*, *IGF1R*, *CCNE1*, and *CDKN2A/B*, and high frequency of deletion of DNA repair-related genes such as *BRCA2*, *PTEN*, *MDM2*, *RB1* and *TP53*). Possible therapeutic drugs for this subtype include poly (ADP-ribose) polymerase (PARP) inhibitors and genotoxic agents (BL1 patients are sensitive to cisplatin treatment[22]).

*Commonly understood as the absence of residual invasive carcinoma in the complete resected breast specimen and all sampled regional lymph nodes[20].

Table 1.1: Genomic TNBC subtypes and potential therapeutic targets, adapted from Collignon *et al.*[5], data from Lehmann *et al.*[21]

Subtypes	Genetic abnormalities	Potential therapeutic targets
BL1	Cell cycle gene expression DNA repair genes (ATR–BRCA pathway) Proliferation genes	PARP inhibitors Genotoxic agents
BL2	Growth factor signaling pathways (EGFR, MET, NGF, Wnt/β-catenin, IGF-1R) Glycolysis, gluconeogenesis Expression of myoepithelial markers	mTOR inhibitors Growth-factor inhibitors
M	Cell motility Cell differentiation Growth factor signaling EMT	mTOR inhibitors EMT- and CSC-targeted treatment
MSL	Similar to M+ Low proliferation Angiogenesis genes	PI3K inhibitors Antiangiogenic therapy Src antagonist
IM	Immune cell processes (CTLA4, IL12, IL7 pathways, antigen processing/presentation) Gene signature for medullary BC (rare TNBC with a favorable prognosis)	PD1/PDL1 inhibitors
LAR	Androgen receptor gene Luminal gene expression pattern Molecular apocrine subtype	Antiandrogen therapy

The BL2 subtype has abnormal activation of signaling pathways such as the EGFR, MET, NGF, Wnt/β-catenin, and IGF-1R pathways. Potential targeted therapies include mTOR inhibitors and growth factor inhibitors such as lapatinib, gefitinib and cetuximab[22].

The M subtype has highly activated cell migration-related signaling pathways (regulated by actin), extracellular matrix-receptor interaction pathways, and differentiation pathways (Wnt pathway, anaplastic lymphoma kinase pathway, transforming growth factor (TGF)-β signaling) and is therefore also called metaplastic breast cancer.[22] The M subtype has sarcoma-like or squamous epithelial cell-like tissue characteristics and is prone to develop resistance to chemotherapeutic drugs. Therefore, M-subtype patients might be treated with mTOR inhibitors or drugs

targeting epithelial-mesenchymal transition[8].

Compared with the M subtype, the MSL subtype expresses low levels of cell proliferation-related genes and high levels of stemness-related genes, HOX genes, and mesenchymal stem cell-specific markers. It is speculated that the MSL subtype patient may be treated with PI3K inhibitors, Src antagonists or antiangiogenic drugs. Studies have reported that dasatinib, an Abl/Src inhibitor, can be used for the treatment of patients with M and MSL breast cancers[22].

The IM subtype has significantly enriched immune cell-associated genes and signal transduction pathways, such as the Th1/Th2 pathway, NK cell pathway, B cell receptor signaling pathway, dendritic cell (DC) pathway, T-cell receptor signaling, interleukin (IL)-12 pathway, and IL-7 pathway. Thus, the IM subtype is highly similar to medullary carcinoma of the breast[8]. It is recommended to use PD1, PDL1, CTLA-4 and other immune checkpoint inhibitors for the treatment of patients with IM subtype breast cancer[22].

The LAR subtype has a significantly different gene expression profile to that of other TNBC subtypes. Although the LAR subtype does not express ER receptor, it presents highly activated hormonal-related signaling pathways (including steroid synthesis, porphyrin metabolism, and androgen/oestrogen metabolism). It is worth noting that the androgen receptor (AR) is highly expressed in the LAR subtype of breast cancer, and its mRNA level is nine times that of other TNBC subtypes. IHC also detects high expression of AR and a large number of downstream metabolic markers of AR and their auxiliary activators in the LAR subtype. Therefore, anti-AR therapy is recommended for patients with LAR-subtype breast cancer.

Through PAM50 (a tumour profiling test), Lehmann & Pietenpol [22] found that, other than the LAR and MSL subtypes, all TNBC subtypes were mainly composed of the BL subtypes (BL1 [99%], BL2 [95%], IM [84%], and M [97%]). LAR subtypes included HER2 (74%) and luminal B (14%); MSL subtypes included basal-like (50%), normal-like (28%), and luminal B (14%).

Masuda *et al.* [23] found that the LAR subtype had a higher distant metastasis-free survival rate and overall survival rate, while those of the M and BL2 subtypes were

poorer. The 3-year recurrence rates of the M and BL2 subtypes were significantly higher than that of the LAR subtype. Furthermore, although the overall pathologic complete response was 28%, there were significant differences in the specific responses between subtypes. The pCR rate of the BL1 subtype was the highest (52%), while those of the BL2, LAR, and MSL subtypes were 0%, 10%, and 23%, respectively.

In 2015, Burstein *et al.*[24] analysed samples from 198 patients and divided TNBC into four subtypes:

- LAR — expresses AR and cell-surface mucin MUC1.
- M — growth factor receptors such as platelet-derived growth factor receptor α (PDGFRα) and c-Kit receptor.
- BLIS (basal-like immunosuppressed) — expresses the immunosuppressive molecule VTCN1
- BLIA (basal-like immune-activated) — expresses STAT signal transduction molecules and releasing cytokines.

The prognostic analysis showed that disease-free survival (DFS) was in the order BLIA > M > LAR > BLIS(p = 0.019) and disease-specific survival (DSS) was BLIA > M > LAR > BLIS (p =0.07).

If compared with the classification by Lehmann *et al.*[21], Burnstein's LAR and mesenchymal types fall into Lehmann's LAR and MSL.

Liu *et al.*[25] integrated the transcriptome profiles of both messenger RNA (mRNA) and long noncoding RNA (lncRNA) in 165 TNBC samples, through which they arrived at a new definition of TNBC subtypes, which they categorised as IM, LAR, mesenchymal like and BLIS.

In 2016, Lehmann *et al.*[26] refined their earlier subclassification, defining only four groups: BL1, BL2, M and LAR, based on a retrospective analysis of a dataset of clinical trials.

The comprehensive analysis of TNBC can lead to the improved selection of study populations that have the highest probability of responding to specific treatments.

In the future, the evaluation of heterogeneity in TNBC and subtyping may lead to different therapeutic strategies.

1.1.2 Overview of current and novel therapeutic strategies

In this section I will review the array of therapies that are used or under testing for TNBC, with an special focus on the attempts to target the epidermal growth factor receptor (EGFR).

1.1.2.1 Chemotherapy

As discussed at the beginning of this chapter, compared to other types of breast cancer, TNBC has limited treatment options, is prone to recurrence and metastasis, and has a poor prognosis[8]. The main reason for this is that the expression of ER, PR and HER2 are all negative — making the usual endocrine therapies and anti-HER2 agents ineffective. Therefore, chemotherapy has become the main approach for the treatment of TNBC. It is becoming clear that the use of neadjuvant chemotherapy has a significantly higher pCR in TNBC than for hormone receptor-positive breast cancers and can significantly improve the prognosis of TNBC patients[8]. Cytotoxic chemotherapy (taxanes, anthracyclines, capecitabine, ixabepilone) is the standard of care for metastatic TNBC[27]. Despite these advances, the duration of the response to chemotherapy in TNBC tends to be short with a median distant disease-free interval of 11.9 weeks for first line chemotherapy[28]. In general, the results point out that treatment for TNBC remains an unmet need in breast cancer care.

1.1.2.1.1 Platinum-based chemotherapy The use of platinum compounds in metastatic BC was first evaluated decades ago[16], and retrospective analyses have observed increased survival in patients with advanced TNBC after treatment with platinum-based CT[29]. A Phase II study, on a total of 86 patients, showed activity of platinum agents (including cisplatin and carboplatin) in patients with metastatic TNBC, especially in patients with germ-line *BRCA1/2* (response rate was 54% vs 25% in overall patients)[30]. But other studies indicate carboplatin can be effective in TNBC patients regardless of germ-line *BRCA* status. Patients

Table 1.2: Potential therapeutic strategies and targeted drugs used in particular subtypes of triple negative breast cancer, adapted from Yin *et al.*[8]

TNBC subtype	**Therapeutic strategies**	**Therapeutic targeted drugs**
BL1 (basal-like 1)	Inhibit cell proliferation and DNA damage response	**Mitosis inhibitors** (Paclitaxel, Docetaxel, Ixabepilone, Nab-Paclitaxel, Vinorelbine) **Cytostatics** (Cisplatin, Carboplatin, Nedaplatin, Eptaplatin, Oxaliplatin, Lobaplatin, Satraplatin, Mercaptopurine) **PARP inhibitors** (Olaparib, Rucaparib, Talazoparib, Niraparib) **DNA synthetic inhibitors** (Topotecan, Irinotecan, Camptothecin, Doxorubicin, Daunorubicin, Mitomycin)
BL2 (basal-like 2)	Inhibit TP63, EGFR, and MET signaling	**Cytostatics** (Cisplatin, Carboplatin, Nedaplatin, Eptaplatin, Oxaliplatin, Lobaplatin, Satraplatin, Mercaptopurine) **PARP inhibitors** (Olaparib, Rucaparib, Talazoparib and Niraparib) **Growth Factor inhibitors** (Erlotinib, Gefitinib, Afatinib, Osimertinib, Olmutinib, Nazartinib, Avitinib, lapatinib, Cetuximab, Panitumumab, Vandetanib, Bevacizumab, Pertuzumab, Ramucirumab, Trastuzumab, Axitinib, Cabozantinib, Ceritinib, Crizotinib, Lenvatinib, Nilotinib, Pazopanib, Regorafenib, Sorafenib, Sunitinib) mTOR inhibitors (rapamycin, everolimus, RapaLink-1)
IM (immunomodulatory)	Inhibit immune signaling	**Cytostatics** (Cisplatin, Carboplatin, Nedaplatin, Eptaplatin, Oxaliplatin, Lobaplatin, Satraplatin, Mercaptopurine) **PARP inhibitors** (Olaparib, Rucaparib, Talazoparib and Niraparib) **Immune checkpoint inhibitors** (Ipilimumab, Nivolumab)
M (mesenchymal)	Inhibit EMT, Wnt, PI3K, mTOR, Scr, TGFβ, IGF1R, Notch	**Growth Factor inhibitors** (Erlotinib, Gefitinib, Afatinib, Osimertinib, Olmutinib, Nazartinib, Avitinib, lapatinib, Cetuximab, Panitumumab, Vandetanib, Bevacizumab, Pertuzumab, Ramucirumab, Trastuzumab, Axitinib, Cabozantinib, Ceritinib, Crizotinib, Lenvatinib, Nilotinib, Pazopanib, Regorafenib, Sorafenib, Sunitinib) **mTOR inhibitors** (Rapamycin, Everolimus, RapaLink-1) **Scr inhibitors** (Bosutinib, Dasatinib) **PI3K inhibitors** (Idelalisib)
MSL (mesenchymal stem-like)	Inhibit EMT, Wnt, TGFβ, MAPK, Rac, PI3K, mTOR, Scr, PDGF	**Growth Factor inhibitors** (Erlotinib, Gefitinib, Afatinib, Osimertinib, Olmutinib, Nazartinib, Avitinib, lapatinib, Cetuximab, Panitumumab, Vandetanib, Bevacizumab, Pertuzumab, Ramucirumab, Trastuzumab, Axitinib, Cabozantinib, Ceritinib, Crizotinib, Lenvatinib, Nilotinib, Pazopanib, Regorafenib, Sorafenib, Sunitinib) **mTOR inhibitors** (rapamycin, everolimus, RapaLink-1) **PI3K inhibitors** (Idelalisib) **MAPK inhibitors** (Trametinib, Dabrafenib) **Scr inhibitors** (Bosutinib, Dasatinib)
LAR (luminal androgen receptor)	Inhibit AR signaling, FOXA1, and ERBB4 signaling	**Nonsteroidal antiandrogens** (bicalutamide) **mTOR inhibitors** (rapamycin, everolimus, RapaLink-1) **PI3K inhibitors** (Idelalisib)

with early TNBC show similar improved outcomes after cisplatin treatment[5]. Patients who presented higher values in assays for *BRCA*-like (see *BRCA1/2* mutation, "BRCAness" and PARP inhibitors on page 17) genomic instability (homologous recombination deficiency [HRD] large-scale state transition and HRD loss of heterozygosity) also responded better to platinum-based treatments, even in the absence of germ-line mutations[30].

Platinum compounds are not currently recommended for the treatment of early TNBC, and the guideline is to use them only for patients with known *BRCA* mutations, but the available data are conflicting. It has been suggested that the higher chemosensitivity of TNBC tumours in patients with germ-line *BRCA* mutations is what leads to better outcomes, even without platinum compounds[5]. More clinical trials are needed to define the optimal standard systemic therapy for TNBC.

1.1.2.2 Immunotherapy

The tumour microenvironment is a crucial factor for cancer cells survival and invasiveness. Host immunity can either stimulate tumour growth or eradicate the disease — tumour cells are initially rejected by the immune system, then remaining surviving tumour cells persist in a state of dormancy and, after upregulating pro-survival pathways, express factors that promote immune suppresion and angiogenesis[10]. Historically, breast cancer has not been deemed an immunogenic tumour. However, there is increasing evidence indicating the prognostic and predictive values of tumour-infiltrating lymphocites (TILs) in BC, which have a positive association with outcome in both early and advanced stages of TNBC[31–33].

The degree of immune infiltration differs among BC subtypes. TIL levels are significantly higher in TNBC and HER2-positive BC[33–35]. On the other hand, the hormone receptor-positive subtype is associated with the least robust TILs presence. The lymphocyte-predominant BC subtype, which contains high levels of TILs (>50%), is associated with improved disease-free and overall survival and pCR in the neoadjuvant context[36,37].

These findings suggest that immunomodulation could represent a new approach in the treatment of these aggressive BC subtypes[35]. Our current understanding suggests that the immunogenic potential of TNBC (which is the most immunogenic type of BC) is derived at least in part from its genetic instability and high mutation rate[5]. Tumours from patients with TNBC are more likely than other subtypes of BC tumours to exhibit chromosomal instability and potential mutations[38,39].

TNBC is also the BC subtype that is most frequently associated with TILs[10], but only a minority of TNBCs demonstrate a high number of TILs, suggesting that IM therapy could be necessary to promote immunorecognition and increase the adaptive immune infiltrate to therapeutic levels in most patients[5]. Patients that already present high levels of TILs at the time of diagnosis might further benefit from the use of drugs that can enhance antitumoral immune responses[5].

The anti-tumour activity of checkpoint immune inhibitors has been extensively studied in TNBC. Programmed cell death protein 1 (PD-1) is an immune checkpoint receptor that limits the action of T-cell within tissues, playing a crucial role in the tumour immune evasion process. Its two ligands are PD-L1 and PD-L2. Regulation of PD-L1 can occur in response to IFN-gamma action, oncogenic signaling, deletion or by the silencing of PTEN and the subsequent overexpression of the PI3K pathway[40]. Immunotherapy with PD-1 and PD-L1 inhibitors promotes the activation of T-cells, restoring anti-tumour immune activity and resulting in increased survival in selected tumours[10]. PD-L1 is present in 15.8-30% of TNBC cases[40,41]. Therefore, PD-L1 scoring has been used as a strong biomarker and can be measured on tumour or tumour infiltrating cells. Furthermore, its benefit as a part of combination therapy has been proven as a first line therapy for PD-LI+ TNBC patients[42].

The clinical trials performed so far present promising results for the use of immuno-oncology agents in TNBC[15]. The future challenge will be the identification of tumoral immune microenvironments that improve prognosis in BC. For example, in TNBC, controlling tumour growth with conventional chemotherapies in combination with immune checkpoint inhibitors could increase response rates[37].

1.1.2.3 Targeted therapy and biomarkers

A main goal of oncology is to selectively target cancer cells, exploiting their distinct malignant phenotype of uncontrolled cell growth and proliferation[43]. Many cancers are driven by specific oncogenic driver mutations, making them dependent on a single pathway, and therefore sensitive to inhibition of single oncogenes, a concept known as "oncogene addiction"[44]. A well known example of this is the use of gefitinib, an EGFR kinase inhibitor, for the treatment of patients with metastatic non-small cell lung cancer (NSCLC) presenting specific mutations in EGFR[45].

Another effective approach in targeted therapy relies on the concept of synthetic lethality, which refers to the property of a couple or a group of genes, where the loss of a single one of them has no effect on cell viability, but their combined loss induces a lethal phenotype[46]. This can occur as a result of a mutation and/or drug treatment. It can also happen through essentiality inducing, where targeting of one gene makes the other one essential[47]. The best example of this is the remarkable sensitivity of BRCA1/2 mutant tumours to PARP inhibition, which will be discussed below.

Table 1.3: Summary of prognostic biomarkers in TNBC, adapted from da Silva *et al.* [10]. *NA: not applicable; DFS: disease free survival; OS:overall survival.*

Biomarker	% of TNBC with expression/-mutations	Main function	Prognostic significance	Targeted therapies
TP53 gene	75-80% (somatic mutations)	Apoptosis	Low gene expression in *TP53* missense mutations correlate with poor prognosis (worst DFS, but conflicting data).	NA
Ki-67	45-53% (high expression, > 20%)	Cell proliferation	High index and high expression correlate with shorter DFS and OS.	NA
EGFR	13-78%	Cellular growth	Increased expression associates with worst DFS.	Erlotinib, gefitinib, afatinib
c-KIT	50%	Cell transformation and differentiation	Predictor of poor cancer-specific survival in patients with TNBC	Imatinib
VEGF	32%-62%	Angiogenesis	High levels associate with disease progression and metastasis rates.	Bevacizumab
Androgen receptor	10-55%	Cell proliferation and dedifferentiation	Positive expression correlates with higher DFS. May be associated with chemoresistance.	Bicalutamide, enzalutamide, abiraterone
BRCA1 and BRCA2 genes	14-20% (germline mutations)	DNA-double strand break repair	Mutated status correlates with increased DFS.	PARP inhibitors - Olaparib
PD-L1 protein	15-30%	Tumor immune evasion process	High expression correlates with a higher survival rates in trials with checkpoint inhibitors	Immune Checkpoint inhibitor - atezolizumab
Notch pathway	~10%	Cell Proliferation and differentiation	Potential target under development	AL101
PI3-kinase pathway	~25%	Cell Proliferation	Multiple genomic alterations lead to activated PI3-Kinase pathway, including activation in PIK3CA, AKT and mTOR or inactivation in tumor suppressor genes such as PTEN.	PI3K inhibitor - alpelisib AKT Inhibitors – ipatasertib, capivasertib

Despite some targeted drugs being highly effective, a large number of them induce robust initial clinical responses but the cancer eventually relapses due to intrinsic or acquired drug resistance[44]. Intrinsic drug resistance exists prior to the drug treatment, which selects it by killing the sensitive cells[48]. Acquired resistance is a response to drug exposure, which selects the cells that acquire genetic or epigenetic modifications that leads to survival and proliferation[49]. An example of this is the T790M mutation in EGFR, which accounts for resistance to the TKIs gefitinib, erlotinib and afatinib in NSCLC[50].

Non-genomic "adaptative resistance" is another mechanism for drug resistance. It involves the dynamic rewiring of regulatory networks to circumvent the drug's effect[51]. This takes the form of pathways cross-talk and feedback regulatory loops that eventually bypass drug inhibition and rapidly results in resistance[52]. Adaptative kinome reprogramming is frequently associated with resistance to kinase inhibitors. For example, TNBC can bypass the inhibition of MEK1/2 by increasing the expression of PDGFR-β[43].

The existence of these resistance mechanisms leads to the conclusion that monotherapeutic targeted agents are not effective, and that combination therapy taking aim at two or more targets could increase treatment efficacy and prevent drug resistance.

TNBC presents high heterogeneity[17] and therefore it is particularly difficult to discover new therapeutic targets for treatment. Many clinical trials targeting specific receptors are ongoing[17]. TNBC has no prominent driver mutation in oncogenes, most of its somatic mutations are in tumour suppressor genes, such as *TP53*, *RB1* and *PTEN*, which are commonly considered *undruggable* targets[53].

Despite this, several promising targets and biomarkers are discussed below. Biomarkers are defined as reproducibly quantifiable biological variables[10]. In clinical oncology practice, they can be measured as parameters to predict survival, or the response to a therapeutic intervention[10] (Figure 1.1, Table 1.3).

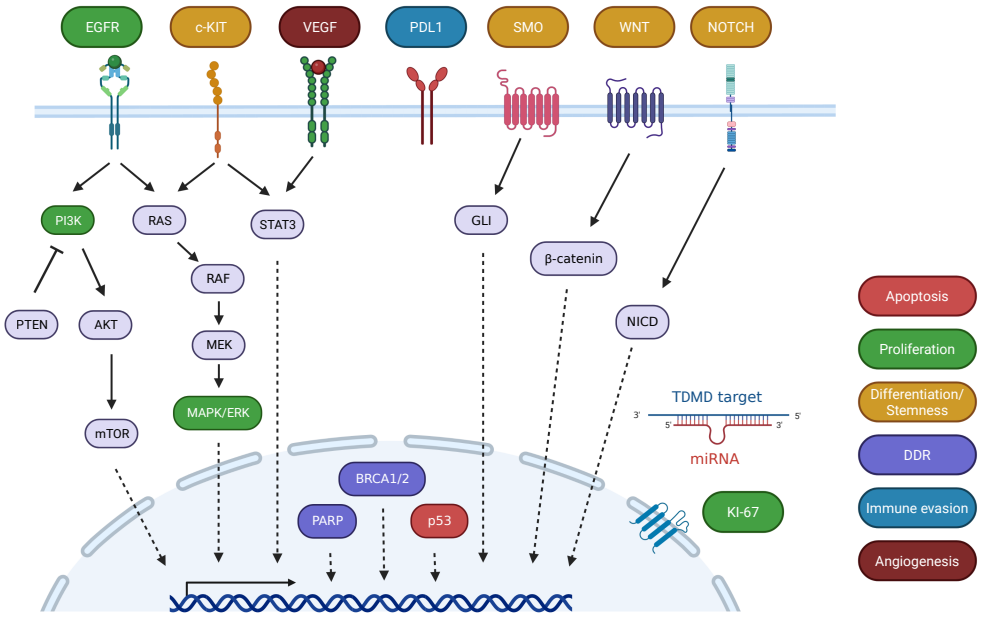


Figure 1.1: Key signal transduction mechanisms in TNBC tumorigenesis. The key proteins are coloured by their function, as described in da Silva *et al.*[10]. Created with BioRender.com

1.1.2.3.1 ***TP53*** The tumor protein 53 gene (*TP53*) encodes the p53 protein, a transcription factor that plays an essential role in the response to DNA damage, acting as a tumour supressor. Transcription of p53 increases when DNA damage occurs, causing cell cycle arrest and DNA repair or cell death. It does so through interaction with targets such as p21, cyclin dependent kinases (CDKs), repair proteins such as PARP, BRCA1, BRCA2, and PTEN[54].

TP53 mutation occurs in approximately 30% of cases of breast cancer and in 75%-80% of TNBCs[55]. The mutant p53 protein may present aberrant interactions with its downstream targets — upregulating CDK-1 and the PI3K/AKT/mTOR pathway, for example, or downregulating tumour suppressing proteins such as p63 and p73, or stimulating cell proliferation and apoptosis evasion[56].

Several studies have attempted to determine the impact of *TP53* mutations on TNBC patient prognosis[10]. As an option for personalised therapies, p53 used to be regarded as *undruggable*, however there are recent methods to selectively target the p53 mutant protein and restore its wild-type properties in breast cancer cells,

such as PRIMA-1 and PRIMA-1^{Met} (p53 reactivation and induction of massive apoptosis 1)[57]. This is a promising targeted therapy option for evaluation in larger clinical TNBC patient cohorts.

1.1.2.3.2 *BRCA1/2* mutation, "BRCAness" and PARP inhibitors BRCA1 and BRCA2 are crucial for homologous recombination repair (HRR), a mechanism to repair double-strand breaks (DSBs)[58]. BRCA1/2 are tumour suppressor genes, since germline mutations in them confer significantly elevated risks of breast, ovarian and other cancers[59].

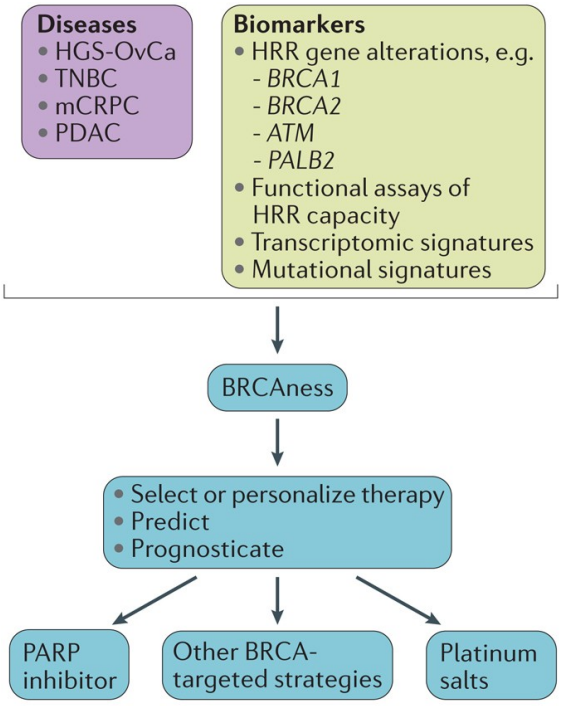


Figure 1.2: Biomarkers and cancer histologies associated with *BRCAness* and its potential clinical utility. From Lord & Ashworth [60]. Some or all of these biomarkers are relatively enriched in cancer histologies such as high-grade serous ovarian cancer (HGS-OvCa), triple-negative breast cancer (TNBC), metastatic castration resistant prostate cancer (mCRPC) and pancreatic ductal adenocarcinoma (PDAC).

BRCA-mutant tumours often harbour mutations in *TP53* and amplification of the *MYC* proto-oncogene, extreme levels of genomic instability, histological characteristics such as pushing murgins and lympocytic infiltrates, and tend to have a TNBC phenotype[60]. Germline *BRCA1/2* are present in approximately 14-20% of TNBCs, but a larger proportion of patients have been reported to harbour homologous recombination deficiency (HRD)[18,19]. HRD can occur in sporadic cancers through genetic and epigenetic inactivation of other genes responsible for HRR, a condition defined as "*BRCAness*"[60,61]. *BRCAness* is usually defined as a phenocopy of *BRCA1/2* mutation, this is, the presence of HRD without a germline *BRCA1* or *BRCA2* mutation[60]. This includes germline or

somatic mutations in genes involved in HRR, such as *ATM*, *ATR*, *CHEK2*, the *FANC* family, *PALB2* or *RAD51*[60].

Importantly, cancers presenting *BRCAness* are susceptible to treatments that target HRR defects, mainly PARP inhibitors, or DNA-damaging agents, such as platinum salts. When HRD is present non-conservative forms of DNA repair, such as non homologous end joining and base-excision repair, are used to repair DSBs[62]. These pathways require PARP1, and because of this its inhibition induces a lethal phenotype in cancers presenting *BRCAness*. In fact, some authors have proposed a broader definition that includes defects in the replication fork protection and other mechanisms that confer synthetic lethality with PARPis[63].

Several studies have attempted to find a biomarker of HRD with the aim of better predicting responders to PARP inhibitors and DNA-damaging chemotherapies.[64,65] (figure 1.2). Experimental biomarkers of *BRCAness* include genetic or epigenetic alterations in genes that control HRR, transcriptomic signatures associated with HRR gene defects, mutational signatures present in *BRCA1* or *BRCA2*-mutant cancers as well as functional assays of HRR function, such as the detection of nuclear RAD51 foci[66] after DNA damage. A well defined *BRCAness* phenotype could be used clinically, either as a prognostic tool or as a way to inform patient selection for therapy.

Early phase II studies of the PARP inhibitor olaparib in BRCA mutation carriers showed encouraging results[67–69], particularly when compared with chemotherapy in patients with germline BRCA mutations, as shown in phase III randomised trials[70]. Another study demonstrated similar results for the PARP inhibitor talazoparib[71,72]. In 2023, the UK National Institute for Health and Care Excellence (NICE) started recommending olaparib as a neoadjuvant therapy for HER2-negative early breart cancer patients with BRCA1/2 mutations. Despite these promising results, resistance to PARP inhibitors remains a problem to be solved[73].

PARPi resistance is very commonly observed in the clinic. More than 40% of patients with BRCA1/2 mutations do not respond to the therapy, and prolongued administration also produces resistance. The main mechanisms of resistance

are the restoration of HRR, the DNA replication fork protection, and reversion mutations[60,74]. The restoration of HRR happens through the action of proteins that control DNA resection and therefore commitment to HRR. This includes cell cycle proteins such as CDKs, and 53BP1 and associated proteins (REV7, RIF1, SHLD proteins), which block end resection by inhibiting BRCA1 recruitment[75–77]. Several studies have shown that CDK inhibitors can reverse PARPi resistance[74,78,79].

PARPi inhibitors also have the ability to trap PARP at the replication fork. When this happens, cells usually rely on *BRCA1/2* to stabilise these stalled forks and prevent them from being degraded. Without this protection, PARP trapping leads to genomic instability and cell death. Several mechanisms that protect stalled forks from degradation in absence of BRCA1/2 may be behind some forms of PARPi resistance[80–82].

Lastly, PARPi resistance can arise from secondary mutations in *BRCA1* or *BRCA2*. These mutations restore the open reading frame of mutant alelles and therefore restablish BRCA function[83,84].

Another promising strategy, with minimal additional toxicity expected, is combining PARP inhibitors with immunotherapy[85]. It may result in enhanced clinical activity given the greater genomic instability in BRCA mutated cancers[15].

1.1.2.3.3 Ki-67 A protein encoded by the *MKI67* gene (marker of proliferation Ki-67 gene), Ki-67 is the most commonly used cell proliferation marker, and it is therefore considered a prognostic biomarker in BC, although with uncertain value when it comes to patients with TNBC[10]. These tumours tend to have increased expression of Ki-67 — studies show a prevalence of 44.7%-53.4% of tumours with positive Ki-67 expression[86].

Recent studies have shown that a higher Ki-67 score is an independent risk factor for disease-free survival and overall survival, and Ki-67 is significantly correlated with the TNBC phenotype[10].

1.1.2.3.4 Androgen receptor The androgen receptor (AR) is part of a complex of steroidal hormone receptors that modulate transcription factors[10]. AR can either promote proliferation and dedifferentiation or induce apoptosis, depending on other signalling pathways. The AR drives the luminal gene expression that characterises the LAR BC subtype[21], though it is expressed in normal and malignant breast tissue, and its prevalence is variable depending on the subtype of BC. Early studies indicated that AR expression in TNBC was linked with worse prognosis, but the latest data have contradicted this[10]. The level of expression of AR in TNBC (measured by IHC) varies widely among cohorts, between 10-90%[87]. Approximately 10%-15% of TNBCs express the AR[88].

Wang *et al.*[89] performed a systematic review of 13 clinical trials with a total of 2826 TNBC patients. 24.4% or the cases were AR-positive, which was associated with low tumour grade, post-menopausal status and lower risk of nodal involvement. Other studies have corroborated this[90], but none directly linked AR expression with overall survival.

There is evidence that patients with AR-positive tumours are more likely to be chemo-resistant than AR-negative patients, with a lower pCR among the former[91,92].

Several clinical trials have evaluated the efficacy and safety of anti-androgens in advanced TNBC, with promising results[5,10]. More data are needed before it can be considered as a new validated treatment option. Future challenges related to the LAR subtype include understanding the role of the AR in tumorigenesis and the escape mechanisms in AR therapy, and discovering predictive biomarkers[5].

1.1.2.3.5 c-KIT (CD117) c-KIT seems to be an important part of cell transformation and differentiation[10]. Its expression is detected in approximately half of TNBC patients, and the aberrant activation of the *c-KIT* gene is part of the process of carcinogenesis and metastatic mechanisms of various human malignancies[10]. In a study of 58 TNBC patients, positive IHC of c-KIT was associated with *TP53* missense mutations, vascular invasion, recurrence and higher Ki-67 proliferation

index[93]. The joint expression of c-KIT and these missense mutations of *TP53* proved to be an independent prognostic factor for worse survival[93].

1.1.2.3.6 microRNAs and long non-coding RNAs Noncoding RNAs are functional RNA molecules, transcribed from DNA but not translated into proteins. They include microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and others.

miRNsA are small noncoding RNA molecules of 17-27 nucleotides in length, which expression signatures have been studied for both early diagnosis and treatment monitoring, as they could constitute reliable early-stage blood biomarkers in TNBC[10]. They have a gene regulatory function through the stimulation or degradation of mRNA targets.

Recent studies show the potential of specific miRNAs as prognostic biomarkers in TNBC[94,95]. The miR-34 family are among the most studied biomarker miRNAs in TNBC[96]. miR-34a is associated with reduced tumour growth and miR-34c with worse prognosis[97]. Several studies have shown that miR-34a acts as a tumour suppressor, inducing cell cycle arrest, apoptosis and senescence in cancer cell lines[10].

lncRNAs are non-coding transcripts with lengths exceeding 200 nucleotides. They modulate transcription of genes by protein regulation and controlling RNA maturation and transport[98]. Augoff *et al.*[99] discovered that a set of lncRNA are consistently aberrantly expressed in TNBC, and may be involved in carcinogenic process and progression. Long non-coding RNA in non-homologous end joining pathway 1 (LINP1) is a lncRNA overexpressed in TNBC. LINP1 enhances dsDNA break repair. Sakthianandeswaren *et al.*[100] showed that blocking it may increase TNBC sensitivity to radiotherapy. Another lncRNA, HULC (highly up-regulated in liver cancer) is also upregulated in TNBC tissues and cell lines and is linked to poorer clinical outcomes, so it could be another therapeutic target in TNBC[101].

1.1.2.3.7 PI3K/AKT/mTOR pathway This pathway is essential in proliferation, differentiation and survival[102]. When deregulated, these proteins play a crucial role in tumorigenesis and in causing resistance to therapies[102]. The pathway proteins may be aberrantly expressed in breast cancer, and somatic mutations may

occur in up to 25% of cases[102]. PTEN, INPP4B and other phosphatases are key regulators of this pathway[103]. Genomic alterations can result in an hyperactivated PI3K pathway, such as activation of oncogenes *PIK3CA*, *AKT* and *MTOR*, or inactivation of tumor suppressor genes such as the aforementioned PTEN[104].

Different subtypes of TNBC present different PI3K pathway mutations. PIK3CA and AKT1 mutations are more frequent in AR positive TNBC, and INPP4B is often inactivated in the basal-like subtype. Several phase II studies have targeted the AKT pathway in TNBC, with encouraging results[42,105,106].

1.1.2.3.8 Vascular endothelial growth factor (VEGF) VEGF mediates the signaling of angiogenesis, which is crucial in the process of tumour growth and spreading[107]. The regulator of its gene expression are hypoxia, growth factors, nitric oxide, oncogenes, HER2 and tumour suppressor genes[10]. 30-60% of TNBC patients present a high expression of VEGF[108]. VEGF and the microvascular density scoring by IHC have been used as a prognostic biomarker in TNBC. High mean vascular density in breast cancer is associated with very poor prognosis and decreased survival[109].

VEGF has been widely studied as a target for treatment of TNBC in different trials of bevacizumab, monoclonal antibody of VEGF-A, with generally disappointing results[5,10]. This includes studies in a adjuvant setting, where bevacizumab did not show any benefit in survival outcome[110], and neither did anti-VEGF tyrosine kinase inhibitors (TKIs)[107].

1.1.2.3.9 Epidermal growth factor receptor (EGFR) A considerable portion of this thesis is focused on the analysis of a dataset coming from a study on a TNBC exceptional responder to EGFR therapy[1], which will be described at length in chapter 2. Because of this, the epidermal growth factor receptor and related pathways and targets, are crucial topics in this thesis, and therefore this section will examine in depth the current landscape of EGFR targeting in TNBC.

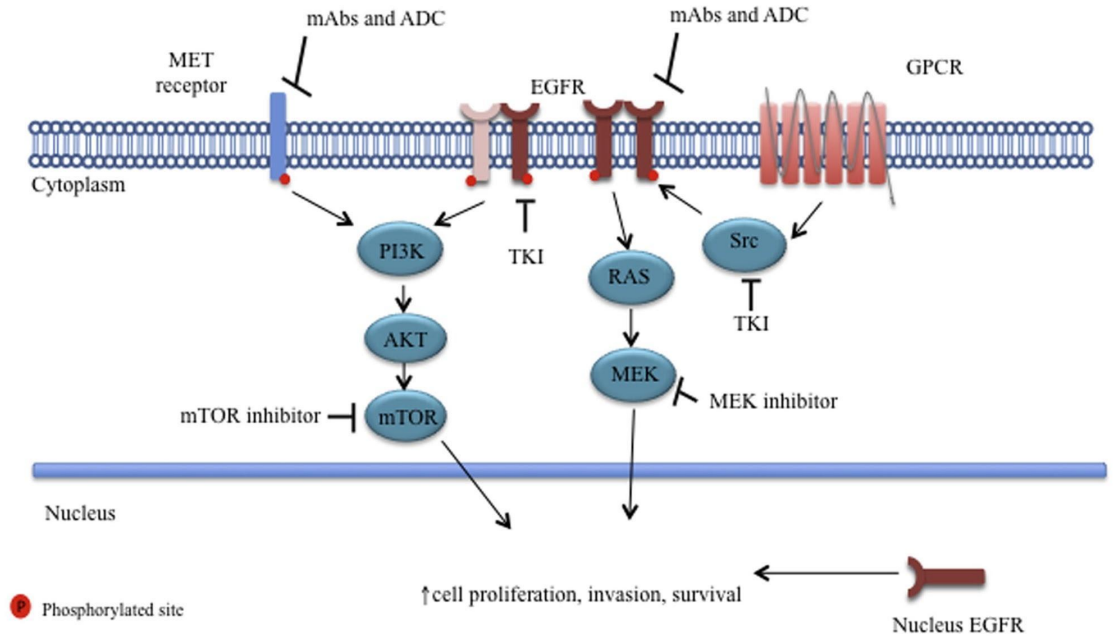


Figure 1.3: EGFR signal transduction. From Costa *et al.*[27]. ADC: antibody drug conjugate; EGFR: epidermal growth factor receptor; GPCR: G-Protein coupled receptor; HER: human epidermal receptor; mAbs: monoclonal antibodies; TKI: tyrosine kinase inhibitor

1.1.2.3.9.1 Biology EGFR is part of the ErbB family of receptor tyrosine kinases (RTKs)[111], a group of transmembrane glycoproteins with a tyrosine kinase domain that activate signal transduction pathways, playing an important role in cell proliferation, differentiation, and apoptosis inhibition[21,112]. It is frequently mutated or overexpressed in several cancer types[111]. The activation of these receptor kinases occurs chiefly via ligand mediated hetero- or homo-dimerization.

When unstimulated, EGFR is in an auto-inhibited, dimerization-incompetent state at the cellular membrane[111]. Ligand binding promotes the receptor dimerization, which causes structural rearrangements, conveyed to the cytoplasmic domain, resulting in the formation of asymmetric dimers in the catalytic domains[113]. This leads to the allosteric activation of the EGFR kinase and the trans-autophosphorylation of tyrosine residues in the cytoplasmic tail, which triggers the signaling cascade[113]. This ligand-dependent phosphorylation also causes the receptor endocytosis - the ligand-bound EGFR has internalisation rates 10-fold higher than unliganded EGFR[111].

EGFR can heterodimerize with other ErbB family members, such as ErbB2/HER2, ErbB3/HER3 and ErbB4/HER4[113]. These heterodimers have a reduced ligand-binding strength, which leads to ligand dissociation in endosomes, and renders them unable to recruit the endocytic machinery as efficiently as EGFR homodimers[114]. Thus, signaling from heterodimers is enhanced and thought to be more oncogenic than that from homodimers[111].

In breast cancer, HER2, the preferred dimerization partner for EGFR, amplifies the signal generated through the dimer receptor complex[27]. Alternatively, cytoplasmic EGFR can also be activated by its extracellular EGF ligand, or by cytoplasmic tyrosine phosphorylation by substrates of G-protein coupled receptor[115,116] (figure 1.3). EGFR then initiates downstream signaling through the PI3K-AKT-mTOR and RAS-MEK pathways, promoting cell proliferation and survival[117]. In addition to this function, EGFR can also act as a cotranscription factor in the nucleus, promoting cancer progression[118]. Nuclear EGFR expression correlates *in vitro* with resistance to the oral tyrosine kinase inhibitor (TKI) gefitinib[119]. It also correlates with a more aggressive clinical behaviour in BC[120].

EGFR is highly expressed in TNBC[121]. Gene amplification has been observed in 6-21% of patients, dependent on the applied method, and overall overexpression is reported as 30-50%, with a significant correlation with poor overall survival[121,122]. Activating mutations are rare, but have been identified in 11.5% of TNBC tumours from Asian patients[123]. Because of this high expression and EGFR effect on cell proliferation and motility, EGFR is regarded as an attractive therapeutic target for TNBC[124–126].

Attending to Lehmann *et al.* [21] classification, EGFR is particularly highly expressed in the BL and MSL subtypes[127]. MSL is enriched for the epithelial-mesenchymal transition (EMT), and growth factor pathways including EGF, PDGF, ERK1/2, G-protein coupled receptor (GPCR) and calcium signaling pathways[21]. The BL2 subtype displays gene ontologies involving growth factors (EGF, NGF, MET, Wnt/β-catenin, and IGF1R) and growth factor receptors (EGFR, MET receptor and EPAH2)[21]. Several of these receptors and downstream molecules have been

implicated in the EGFR signaling pathway. The Src oncoprotein family participates in the GCPR-induced EGFR transactivation in tumour cells, and PI3K/mTOR is downstream to EGFR in the oncogenic transmembrane signaling cascade[115,128]. MSL cell lines are more sensitive to the Src inhibitor datasinib, and to the Pi3K/mTOR inhibitor NVP-BEZ235 than other subtypes. This supports the hypothesis that the biology of a subset of TNBC is driven by growth factor receptor signaling pathways, including EGFR[129].

1.1.2.3.9.2 Anti-EGFR therapy Several EGFR antagonists have been approved for clinical use in other cancers, such as non-small cell lung cancer and colorectal cancer[126]. Nevertheless, using anti-EGFR as a monotherapy has marginal efficacy in clinical trials in TNBC due to drug resistance[130,131]. In spite of the EGFR-targeted therapies efficacy in solid tumours (colorectal, head and neck, and non-small cell lung cancer)[132–134], the clinical experimental data has showed that EGFR-targeted treatment for TNBC does not achieve the expected results. The results of EGFR signaling pathway analysis in TNBC patients treated with EGFR-targeted therapies showed that the EGFR downstream signaling pathways were still activated in most patients after EGFR-targeted treatment, suggesting that there might be other pathways involved in a bypass activation[126]. As a result, EGFR-targeted treatment alone cannot achieve significant efficacy in TNBC. Nevertheless, there is very active research on this matter, and I will provide an overview of the current landscape.

Resistance to EGFR monotherapy is not restricted to TNBC — it has been reported for many other cancers. EGFR resistance is often associated with overexpression of MET or its ligand HGF[135,136], co-inhibition of EGFR and MET in TNBC models had synergistic effects *in vitro* and tumour growth inhibition in animal models[137,138]. The patients with TNBC coexpressing *MET* and *EGFR* had inferior disease-free survival compared to those expressing *EGFR* alone[139].

The RTK AXL is also up-regulated in response to EGFR antagonists and participates in EGFR-targeting resistance in lung cancer[140]. AXL is highly expressed in the TNBC M subtype and desensitises tumour cells to EGFR inhibition[141].

MEK is another target to be considered, as there is evidence of high levels of activation of the RAS/MEK/ERK pathway in TNBC[142]. Different studies have shown the synergistic effects of combining the EGFR inhibitor gefitinib with MEK inhibition[143,144].

EGFR is an upstream regulator of the aforementioned PI3K/AKT/mTOR pathway[115]. Liu *et al.*[145] examined the effects of combining EGFR inhibition (with lapatinib) and mTOR inhibition (with rapamycin) *in vitro* and *in vivo*. This resulted in a synergistic significantly greater cytotoxicity than that of any of the single agents alone. *In vivo*, it also significantly decreased tumour growth in cell derived xenografts[145].

Tao *et al.*[146] showed that inhibition of EGFR in TNBC provokes HER3 up-regulation — HER3 downstream signalling compensates for EGFR inhibition through the loss of AKT signalling and an AKT mediated negative feedback loop. They proposed that the simultaneous blockade of EGFR, HER3 and the PI3K-AKT pathway could be beneficial in TNBC.

Another strategy to circumvent HER3-associated resistance to EGFR is co-targeting of EGFR and PYK2[147], a non-receptor tyrosine kinase. PYK2 is a downstream effector of EGFR and their joint high expression in TNBC correlates with overall reduced survival in basal-like tumours[43,147]. The exact mechanism through which this happens was studied using mathematical modelling, which suggested that PYK2 inhibition facilitates the proteasomal degradation of HER3[148]. Co-targeting of these two molecules inhibited cell growth *in vitro* and tumour growth in the animal model. Moreover, it circumvents HER3-associated resistance[43]. The model predicted that this combination of EGFR and PYK2 targeting eliminates negative network rewiring and reactivation of ERK and STAT3 caused by inhibition of either EGFR or PYK2 alone and this was validated experimentally[149]. The model also performed stratification of TNBC patients to detect subgroups that may be responsive to this combination therapy[149]. This study is one example

of computational modelling being a powerful approach to predict drug-induce resistance and synergistic combinations[150,151] (see section 1.2.3).

Several TKIs have been tested in TNBC, such as lapatinib[152], afatinib[153], or neratinib[154] with disappointing results. Despite ongoing efforts to re-purpose existing TKIs for the treatment of TNBC (such as erlotinib), it seems unlikely that TKI only treatment of patients with unselected TNBC will show significant activity[27,155].

Other studies sought to achieve synthetic lethality combining EGFR and PARP inhibition. Nowsheen *et al.* [156] combined lapatinib with ABT-888 (a PARP inhibitor) in TNBC and achieved synthetic lethality *in vitro* and *in vivo*. This combination significantly reduced the TNBC cell survival fraction in a dose dependent manner by 70-99%, while lapatinib alone only produced a 10-30% reduction. Similarly, neither ABT-888 nor lapatinib alone significantly delayed tumour growth in cell line MDA-MB-231 xenografts, but the combination of both did[156]. More research is needed to confirm the efficacy of this strategy in the clinical setting.

The relationship between immunotherapy and EGFR targeting has also being studied. There is some evidence of interaction between immune checkpoint inhibition and the EGFR pathway, such a correlation between amplification of PD-L1 and PD-2, and of EGFR[157,158] or between TIL count and EGFR-directed mAb treatment[159], but these interactions need to be better understood for drug development in TNBC.

1.1.2.3.9.3 Monoclonal antibodies EGFR targeted antibodies are another potential avenue for treatment. Monoclonal antibodies (mAbs) target cancer cells by attaching to cell surface antigens. Targeting EGFR, they can work through direct tumour cell death (via inhibition of cell survival signaling pathways or ligand-receptor blockade and dimerization inhibition)[160,161]; stimulation of cytotoxic tumour killing cells[162–164]; or through antibody-dependent cellular cytotoxicity (ADCC)[165,166].

An example of EGFR-targeted mAb is cetuximab, a chimeric inmunoglobulin (Ig) G1 mAb that binds to EGFR in cancer cells with higher affinity than EGF or TGF-α, thus preventing EGFR phosporylation[167]. Cetuximab effectively inhibits TNBC growth in pre-clinical studies, alone or in combination with chemotherapy

or TKIs[168]. A randomised phase II trial compared the effects of cisplatin plus cetuximab with cisplatin alone[169]. The primary end point was not met, but the addition of cetuximab doubled the overall response rate (ORR) and appeared to prolong progression free survival (PFS) and OS (overall survival), although the difference was not statistically significant. Another phase II trial using carboplatin and cetuximab[170] yielded similar modest results. Panitumumab is a fully humanised IgG2 mAb that competitively inhibits EGFR binding, which has also been tested in phase II clinical trials[171] with a similar outcome to the cetuximab trial.

Clinical studies have hitherto failed to deliver encouraging results in the use of EGFR-directed mAbs in unselected TNBC[27]. Combination monoclonal antibody treatment is one approach that seeks to address this. The underlying hypothesis is that combining monoclonal antibodies that engage different epitopes significantly accelerates receptor degradation[172]. A study showed this approach reduced surface receptor levels by up to 80% in cell lines[173]. The efficacy of cetuximab on a xenograft model was enhanced by combination EGFR-mAbs treatment in a study by Ferraro *et al.*[174].

Antibody drug conjugates (ADCs) are a rapidly growing class of cancer drugs that combine the targeting properties of mAbs with the antitumour effects of cytotoxic treatment[175]. ADCs constitute an attractive option for treatment as they do not depend on downstream signaling for efficacy but rather on target expression. One such ADC, ABT-414, combines a potent microtubule inhibitor (MMAF) and the EGFR-targeted humanised mAb ABS-806[176]. ABT-414 selectively kills tumour cells overexpressing wild-type or mutant forms of EGFR. It inhibits the growth of patient derived xenografts with high EGFR expression and causes complete tumour regression and cures in the most sensitive models[177].

1.1.2.3.9.4 Conclusions A large number of clinical trials investigating EGFR related therapies for TNBC have been carried out in the last years, for a in depth summary of these, see the reviews by Nakai *et al.*[126], Collignon *et al.*[5], and Bai *et al.*[178]. Overall, EGFR does not perform well as a single therapy, but it remains an

intriguing target in TNBC, due to the studies that give some indication of its efficacy as part of a combination therapy or for a stratified subset of patients[1,27,173,176,177]. As discussed, targeted therapies based on a single agent are not effective, possibly due to the rapid emergence of drug resistance[43]. Combination therapy is the key to overcome resistance and to sustain drug efficacy over time, especially if the drug targets display synthetic lethality. Computational modelling, which will be discussed in the next section, can be a powerful tool to find the right combinations of targets and to predict patient response[43].

1.2 Mathematical modelling in oncology

The last couple of decades have seen a vast outspread of mathematical modelling in oncology. The number of publications on mathematical oncology grows at an exponential rate[179], and its use is recognised as an invaluable tool for the generation of experimental and clinically testable hypotheses, and to guide study design[180].

A mathematical model is a distillation of biological complexity into a manageable representation, which is both dense enough to keep the general properties of the real system and simple enough to yield actionable insights that can be used to make predictions[181]. Mathematical models can help us understand complex biological systems by providing approximations and abstractions that can generate testable hypotheses on how biological components evolve and interact with each other to produce a particular outcome (Figure 1.4).

Modelling can provide greater understanding of the fundamental mechanisms regulating biological systems, and deepen understanding of disease. Recent advances in high-throughput technology have enabled the acquisition of an increasing amount of data in several biological systems, including whole genome sequencing data in disease models and samples from clinical trials[182–186]. One can also build predictive models of therapy outcome using genomic data collected from thousands of clinical samples, or investigate how signalling networks work using functional data from experiments modulating the expression of thousands of genes[187].

It is clear that cancer is a heterogeneous disease, and that different cancers present different genetic, transcriptomic and epigenetic footprints with important implications on their evolution and progression, and therefore on clinical recommendations[17]. It is increasingly evident that one size does not fit all and instead, each type of tumour should be treated in a way that integrates its biological characteristics. This is the goal of "precision medicine"[188], which aims at stratifying patients into coherent groups based on their likelihood to respond to a given therapy, in order to achieve the most effective treatment and minimal risks[189].

Stratification has been pursued in the context of several types of cancer, both in the quest for better drugs to use in combination with existing treatments and in the development of biomarkers[190,191]. Indeed, one of the best examples of stratification is the case of tamoxifen in breast cancer[192]. However, when it comes to anticancer therapies, the number of drugs, their possible combinations and the different existing strategies of dosage scheduling make predicting the patient's response to therapy a truly challenging task, requiring very large clinical datasets and extraction of meaningful and clinically relevant information from them. Mathematical modelling and powerful computational techniques enabling the implementation and development of increasingly complex models can help greatly in this task, by bringing a new sophisticated set of tools to carry out integrated analysis of different data types (e.g. genetic, transcriptional and epigenetic) and experimental systems (*in vitro*, *in vivo* and clinical). High-throughput computing has been employed to explore high dimensional parameter spaces in mechanistic frameworks such as PhysiCell[193,194]. This allows high throughput hypothesis testing, this is, using data driven methods to test hypothesis across hundreds or thousands of simulations[193]. It is becoming increasingly possible, for example, to perform systematic large-scale and multiscale *in silico* experimentation and, once validated models are established, predict a patient cohorts' response to specific treatments[195–197]. As a result, sufficiently validated models could inform decisions such as the combination of drugs to be administered to specific groups of patients, or the dosage that has maximum efficacy and minimal toxicity. This has often been

referred to as "digital twins", models fit with data from patients that continuously integrate observed changes in the patient state and guide their treatment and those of patient cohorts with similar disease states[195–197].

An important issue that must be considered when discussing the role of mathematical modelling in oncology is tumour plasticity. Individual cancer genomes are heterogeneous, which in turn fosters rapid evolution, cancer cell adaptation and treatment resistance through clonal selection[198] and other mechanisms. Genomic profiling on multiple spatially separated cancer samples has shown increasing evidence of clonal evolution, with more than half of all somatic mutations detectable in different subregions of the same tumour[199]. Therefore, there is a need to determine both location and time of driver events, and understand how the prevalence of different subclones over time affects therapeutic response. Mutational landscape is not the only characteristic which changes during treatment. It is well known that a radiotherapy treatment can modify the tumour microenvironment both in terms of physiological characteristics, such as the presence of nutrients or oxygen[200], and the populations of cells present[201]. All this implies that therapies should be adjusted or adapted to the emerging context, instead of using a static approach whereby the treatment is mostly determined upfront[202,203].

The use of imaging and multiple tumour-biopsy samples is delivering an extremely valuable understanding of tumour progression, invasion and response to treatment[204], and could have a great role in therapy modulation. However, these approaches are currently limited by the feasibility and cost of obtaining the necessary data at repeated time points, and thus unlikely to provide the answer for routine clinical practice. Consequently, mathematical modelling and computational tools present a tremendous potential to complement these methods and to reduce the number of data points required to predict the best treatment schedule and the necessary adaptation for each patient.

An exciting new trend in cancer modelling is spatial models. These can mimic the progression from normal tissues all the way to invasive tumours by representing each cell as a three-dimensional (3D) or two-dimensional object. There has recently

been a lot of interest for this kind of approach, and it constitutes a promising avenue in cancer modelling (for a recent review see Karolak *et al.*[189]).

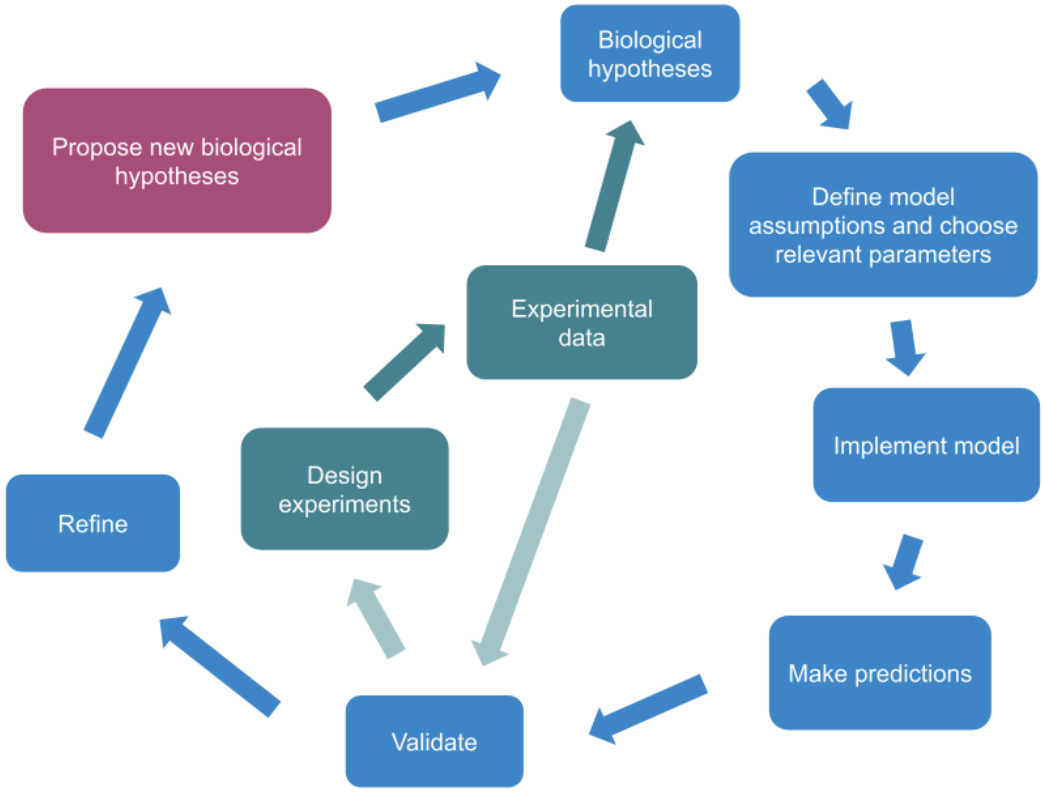


Figure 1.4: A workflow for modelling biological systems. The cycle of designing and refining a model (in blue), experimentally validating it (in green) and then using it to generate new hypotheses (in purple).

1.2.1 Phenomenological and mechanistic modelling approaches

Most sciences present two main modelling philosophies: the mechanistic ("bottom-up") and the phenomenological ("top-down") approaches (Figure 1.5). There is not a clear hard border between the two, and often the same mathematical or statistical methods can be applied to one or the other. However, they differ on how they attempt to explain the relationship between the variables to be modelled.

Mechanistic models are the product of the hypotheses that seek to explain the data obtained, and thus they describe causal relationships between biological components.

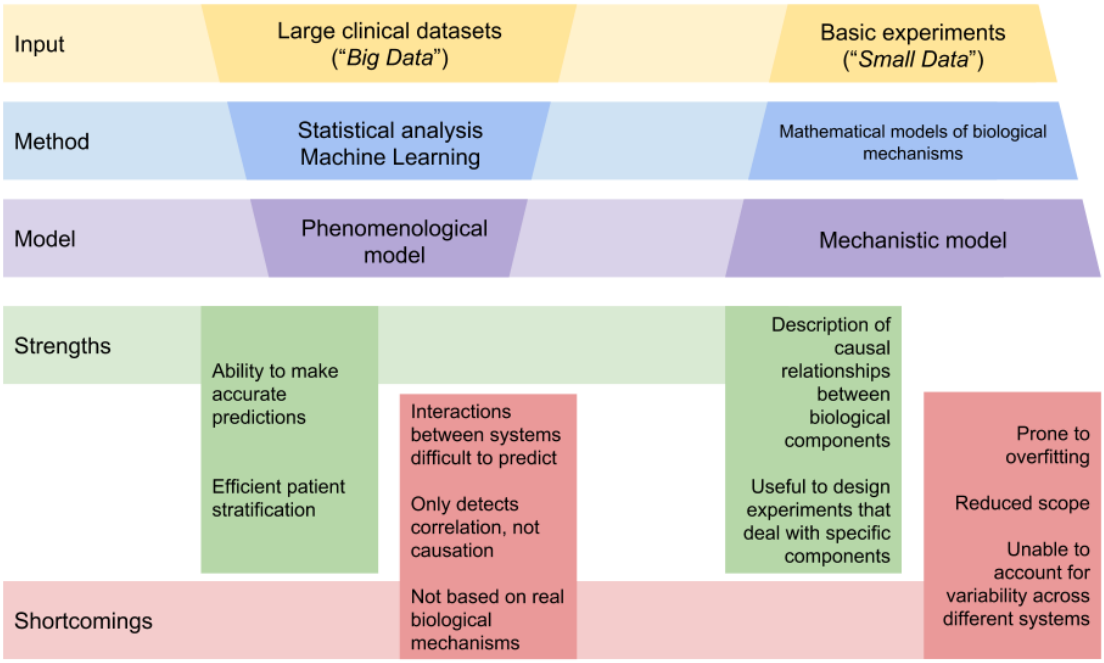


Figure 1.5: Phenomenological and mechanistic approaches to biological modelling, with their strengths and shortcomings.

On the other hand, phenomenological models use observed correlation between variables to predict the outcome of statistically similar experiments[180].

They both have advantages and limitations. Since mechanistic models deal with real biological components, of which there is a vast amount, all presenting complex interactions, they are prone to over complexity and overfitting. Overfitting occurs when a model includes too many variables from a particular set of data and therefore matches it very closely, but is unable to generalise to other data sets and generate reliable predictions for these. Their scope is also more reduced because their components are specific to the system they describe, but this makes them better at informing the design of new experiments that manipulate the specific biological components. Nevertheless, this makes these models excellent tools for the generation of testable hypotheses, which can inform in a very detailed manner the design of new experiments that manipulate those biological components in a precise way. Nevertheless, mechanistic models also need to rely on basic phenomenological data at the level of the interactions between the biological entities they represent. Therefore, the difference between both philosophies is more in where they put the

focus — mechanistic hypotheses vs correlation between observed variables than any fundamentally distinct source of information.

Phenomenological models can serve to make predictions, and they avoid the over-fitting of data by reducing the number of parameters necessary to describe the collective behaviour of the system, but they are not as useful to test specific hypothesis about how a given biological system works. Furthermore, they use correlation, which does not necessarily imply causality or real association, so they can lead to false conclusions. Again, since they don't provide mechanistic explanations, it is more difficult to use them to predict how two different systems will interact.[181].

In many scientific disciplines, these two approaches have been used together in a complementary manner. In cases where the number of variables involved is small, the two approaches can lead quickly to similar results regardless of whether one starts from the hypothesis or from the data. Still, this is an important discussion in biological modelling, since biological systems are incredibly complex entities to model, with hundreds of variables determining their behaviour. While the number of components implicated in, for example, a biochemical pathway or a gene regulatory network is huge (and so is the amount of data scientists have gathered), a model needs to reasonably reduce this to a manageable representation for it to be useful and capable to make predictions. The main problem is that the parameters of the model are inferred from the data, so with both over and under fitting come new problems, and both prevent trustworthy predictions. A good discussion on this topic[181], as well as several examples of both mechanistic and phenomenological models[180,205,206] are provided in the bibliography.

Phenomenological approaches to modelling have seen a wider adoption in medicine as a data-driven approach can often be more effective at conveying the general properties of complex systems[205]. On the other hand, whilst models developed in such a way have been useful to predict the effect of specific therapies in certain patients, they have not provided alternative actionable solutions for those patients who do not respond to those particular therapies. Thus, a parallel mechanistic method would need to complement those models with a biologically led simplification

of the system. This would allow to predict how a system will behave in response to different drugs, and their possible combinations, or after further acquired gene mutations. *microCJ*, the model I have created (presented on chapter 4) is an attempt to bridge both approaches into integrative models[181]. I propose that multiscale spatial models are particularly suitable candidates for this bridging between approaches as they provide a framework in which to simulate a system at several levels of complexity: the intracellular and intercellular interactions, and how the intracellular components can influence the intercellular behaviours and spatial organisation.

1.2.2 Boolean gene regulatory networks

Biological processes are characterized by complex molecular interactions that have been represented using a number of different frameworks. Network representation has been successfully used in various forms. Boolean gene regulatory networks (BGRN) are a powerful such instruments due to their simplicity and computational properties[207]. They are a main component of the research I present in this thesis, and for that reason in this section I will provide a brief explanation of the main theoretical concepts.

These qualitative models allow to distill the highly complex and stochastic nature of biological processes into a model that is simple to understand and interpret. In these networks, all components —be it genes, proteins, regulatory interactions or a mixture of those— are modelled as nodes governed by binary functions, so that at any given time each node has one of two possible states,one or zero (abstracted as on/off, or active/inactive).

They differ from other modelling paradigms (which will be presented in the following sections) in that they are not a model for tumour growth, or spatial organisation, and they are not quantitative: they model the relationship between genotype and phenotype for a single, ideal cell, how initial perturbations in the nodes representing genes can produce different outcomes or behaviours.

More specifically, BGRN are directed graphs, where the relationships between these nodes are expressed as logical Boolean functions. Thus, the state of each node depends on the state of other nodes, according to the specific function for each node.

A Boolean network can be classified according to its update strategy.The three main types will be described in the following sections.

1.2.2.1 Synchronous networks

The most basic case is the **synchronous network** (SBNs)[208], which consists of a set of variables

$$X = \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{B} \tag{1.1}$$

and a set of transition functions

$$F = \{f_1, f_2, ..., f_n\}, f_i\Big(\vec{x}(t)\Big) = x_i(t+1), f_i : \mathbb{B}^n \to \mathbb{B} \tag{1.2}$$

Each function thus maps one Boolean variable in X to a new Boolean value. The state of such a network at time t is defined by the vector

$$\vec{x}(t) = (x_1(t), ..., x_n(t)) \tag{1.3}$$

Here, t represents a discrete update of the network, or "time step". These steps are an abstract measure of how many next states have been computed for a given network, and do not necessarily have a direct correlation with real time, but depending on the model, they can be given one. For this section, I will take t to mean the number of times a network's state has been updated.

A network of size n can present a total of 2^n possible states. The successor state $\vec{x}(t+1)$ is computed by applying the n-dimensional state transition function T_{sync} to $\vec{x}(t)$

$$\vec{x}(t+1) = T_{sync}\Big(\vec{x}(t)\Big) \tag{1.4}$$

with $T_{sync} = (T_1, T_2, ..., T_n)$ and $T_i(x) = f_i(x)$. Therefore all functions are applied simultaneously to reach the successor state $\vec{x}(t+1)$. SBNs are deterministic, to every state $\vec{x}(t)$ there corresponds one and only one successor state $\vec{x}(t+1)$.

1.2.2.2 Asynchronous networks

Asynchronous networks[209,210] consist of a set of variables X and set of transition functions F as above. However, at each time t only one node i^* is chosen at random and updated. $\vec{x}(t+1)$ is computed by applying the transition function $T^{(i^*)}_{async} = \left(T_1^{(i^*)}, T_2^{(i^*)}, ..., T_n^{(i^*)}\right)$ to $\vec{x}(t)$, where

$$T_i^{(i^*)}(x) = \begin{cases} f_i(x) & i = i^* \\ x_i & i \neq i^* \end{cases} \tag{1.5}$$

Therefore, asynchronous networks are non-deterministic, and every state $\vec{x}(t)$ has n possible successor states $\vec{x}(t+1)^{(i^*)}$. The reasoning behind the development of asynchronous updating was that it could be more representative of biological systems in that gene regulation processes do not happen simultaneously for every gene, as it is modelled in a synchronous network. However, since the node to be updated at every transition is chosen at random, the timing of the model is often unrealistic: a biological process can require minutes in a real system, but the set of downstream regulated genes may not be updated for days according to an asynchronous simulation[207]. Moreover, biological processes depend on each other — a protein needs to be transcribed before it can be active, for example. More importantly, different studies have revealed that synchronous and asynchronous networks lead to the same stable biologically meaningful dynamic behaviour[211–213]. And finally, if dealing with a large network, the run time for asynchronous simulation can become a strong limitation[214], although there have been some recent attempts to solve this[215].

1.2.2.3 Probabilistic Boolean networks

Probabilistic Boolean networks (PBNs) were developed to deal with incomplete knowledge about biological processes and with their stochastic nature[216]. In both

of the types explained above it is assumed that all interactions take the same time and/or are equally likely to happen. PBNs can address these limitations[216]. In these networks, each node's behaviour is described by more than one Boolean function. Whenever a node state has to be updated, one of those functions is chosen according to a given probability. In this framework, the outcome of the network —its phenotype— is probabilistically, rather than deterministically, dependent on its previous state. This behaviour more closely resembles stochastic biological processes, while maintaining the underlying simplicity of Boolean networks. Formally, PBNs consist of a set of variables X, same as the other types of network, and to every node $x_i \in X$ there corresponds a set

$$F_i = f_1^{(i)}, f_2^{(i)}, ..., f_k^{(i)} \tag{1.6}$$

where each $f_j^{(i)}$ is a possible transition function for node x_i, with k_i being the number of alternative transition functions for x_i. For every set F_i there is a corresponding set

$$C_i = c_1^{(i)}, c_2^{(i)}, ..., c_j^{(i)} \tag{1.7}$$

$$\sum_{j=1}^{k_i} c_j^{(i)} = 1 \tag{1.8}$$

where every $c_j^{(i)}$ is the probability that function f_j is chosen for node x_i. Therefore, it can be said that a PBN is made up of a different SBN for each possible combination of transition functions for each of its nodes. The total number of possible SBNs that can be constructed from a PBN is

$$L = \prod_{i=1}^{n} k_i \tag{1.9}$$

A state transition is performed by choosing a network $N = (f_{j_1^*}^{(i)}, f_{j_2^*}^{(i)}, ..., f_{j_n^*}^{(i)})$, according to these probabilities and applying a synchronous update. The probability that such a network N is chosen is

$$P_N = \prod_{j=1}^{n} c_{j*}^N \tag{1.10}$$

$$\sum_{N=1}^{L} P_N = 1 \tag{1.11}$$

For each pair of states $y = (y_1, ..., y_n)$ and $z = (z_1, ..., z_n)$, $\mathcal{N}_{yz}$ is the set of networks that yield z as the result of a transition from y. Then, the total probability of a transition from state y to z is

$$\sum_{N \in \mathcal{N}_{yz}} \prod_{j=1}^{n} c_{j*}^N \tag{1.12}$$

1.2.3 Modelling therapies against cancer

Modelling has been applied to a broad range of therapeutic modalities in cancer. In this section, I provide some examples of it.

Many models[180,217,218] have been developed to date to predict the effect of chemotherapy. One major problem in chemotherapy is the delivery of the drug to the cell: ensuring that it arrives in a sufficient quantity and that systemic toxicity is kept in check. A big reason for the failure of many chemotherapeutic agents is the complexity of interstitial drug transport, which is hard to reproduce experimentally. To investigate drug penetration, Rejniak *et al.*[219] developed a model of interstitial transport with the biophysical properties of the tumour tissue that, using the regularised Stokelets method, allowed the simulation of both advective and diffusive interstitial transport. The model showed that if both modes contribute to drug movement equally, the depth of tissue penetration by the drug molecules strongly depends on tissue architecture. This means that a histology sample from a patient could help determine how well a specific drug will perform on a given tumour.

In radiobiology, the use of mathematical modelling has a longstanding tradition, as exemplified by the linear-quadratic (LQ) model[220–222], and is especially useful, because every treatment is made up of a large number of variables (such as total

dose, dose distribution, dose-rate or patient geometry and physiology) which all need to be considered during treatment planning[223].

Modelling of radiotherapy allows researchers, starting from basic assumptions, to estimate the separate contribution of each of these variables and biological parameters to treatment outcome, e.g. local control of the tumour, and the likely risks associated with said treatment. The alternative to this (mechanistic) approach would be a complex series of randomised clinical trials where every change in each of these variables was assessed on its own in a data-driven (phenomenological) manner[222]. This would be prohibitively costly both in terms of time and money, and in most cases unfeasible[224]. These are the reasons why a simple model such as the LQ model, first formulated in 1942[220], is still used today for radiotherapy, either as originally devised or, more often, as the underlying conceptual basis for more evolved and clinically relevant models.

1.2.3.1 Classification of mathematical models

Bull & Byrne [225] recently proposed a classification of mathematical models in six "Hallmarks of Mathematical Oncology", in analogy with the hallmarks of cancer defined by Hanahan & Weinberg [226]. Each hallmark present a decision with respects to the level of biological detail incorporated into the model, the complexity of the model solutions and the behaviours it can reproduce. They are: single versus hybrid frameworks, homogeneity versus heterogeneity, spatially averaged versus spatially resolved, single-scale versus multi-scale, deterministic versus stochastic, and continuum versus discrete[225]. Different oncological models will combine these characteristic in different ways.

1.2.3.2 Ordinary differential equation models

The simplest mathematical models of cancer are those that model solid tumour growth as systems of time-dependent ordinary differential equations (ODEs). These models consider the tumour as a homogeneous population, spatially uniform, and predict how the volume of the tumour will change with time. The most basic such model could be expressed as the exponential law:

$$\frac{dV}{dt} = \lambda V, \ with \ V(t=0) = V_0 \tag{1.13}$$

with solution:

$$V(t) = V_0 e^{\lambda t}.$$

$\lambda < 0$ and V_0 denote the tumour's net growth rate and its initial volume, respectively. There are no growth constraints, levels of nutrients and growth factors are not taken into account, therefore the tumour growth will be unbounded.

However, in reality, tumour growth rate eventually declines and saturates as it increases in size, due to the competition for nutrients and space. To include such competition in the model from equation (1.13), the logistic growth law can be used:

$$\frac{dV}{dt} = \lambda V\left(1 - \frac{V}{K}\right), \ with \ V(t=0) = V_0 > 0 \tag{1.14}$$

with solution:

$$V(t) = \frac{KV_0}{V_0 + (K - V_0)e^{-\lambda t}}$$

where $K > 0$ represents the population's carrying capacity. At large t, the tumour volume $V(t)$ approaches K. Equation (1.14) predicts exponential growth when the tumour is small and growth saturation when $V = K$. To add flexibility to the growth dynamics, the generalised logistic model adds another variable on top of equation (1.14), $\theta > 0$:

$$\frac{dV}{dt} = \frac{\lambda}{\theta} V\left[1 - \left(\frac{V}{K}\right)^{\theta}\right], \ with \ V(t=0) = V_0 \tag{1.15}$$

with solution:

$$V(t) = K\left(\frac{V_0^{\theta}}{V_0^{\theta} + (K^{\theta} - V_0^{\theta})e^{-\lambda t}}\right)^{1/\theta}$$

Equation (1.15) reduces to equation (1.14) when $\theta = 1$.

A large number of growth models have been used to describe that of solid tumours[227], which add further parameters. This can make the model fit experimental

data better, but can also result in overfitting, making the model less general and thus less useful. Some techniques have been proposed in order to tackle this problem, see Liu *et al.* [228] for an example. Furthermore, because these model lack the spatial component, system parameters such as λ and θ are difficult to fit with actual cell behaviour, a weakness adressed by the models in section 1.2.3.3 and section 1.2.3.4.

Such ODE models assume, for simplicity's sake, that the tumour is spatially and phenotypically homogeneous. Again, this is addressed by the models in section 1.2.3.3 and section 1.2.3.4, but ODE models can also account for elements such as vasculature or different subpopulations of cancer cells, but they tend to limit on a small number of such elements. Paczkowski *et al.* [229] use the classical Lotka-Volterra model to analyse the interactions between populations of cancer cells resistant and sensitive to radiotherapy. Their model can be summarised as:

$$\frac{dV_S}{dt} = \lambda_S V_S \left(1 - \frac{V_S}{K_S} - \gamma_R \frac{V_R}{K_S}\right) \tag{1.16}$$

$$\frac{dV_R}{dt} = \lambda_R V_R \left(1 - \frac{V_R}{K_R} - \gamma_S \frac{V_S}{K_R}\right) \tag{1.17}$$

where $V_S(t)$ and $V_R(t)$ denote the volumes of the radiosensitive and radioresistant subpopulations, λ_S and λ_R their growth rates, and K_S and K_R their carrying capacities. If any of the $V = 0$, the other population with grow following the logistic model in equation (1.14). Finally, γ_R and γ_S represent the positive or negative effect that one population has on another's growth. If both are positive, there is competition, both inhibit the growth of the other. On the other hand, if $\gamma_R < 0$ and $\gamma_S < 0$, they are mutualistic, they both enhance the growth of the other, whereas if one is negative and the other is positive, the first one will support the growth of the second, but the second will hinder the first one's growth.

Other authors have employed ODEs to model the impact of vasculature[230,231] or of the immune system[232,233] on growth and radiosensitivity, as well as combination therapy with chemotherapy[234], antiangiogenic therapy[231] or virotherapy[235].

1.2.3.3 Partial differential equation models

Partial differential equation (PDE) models aim to address a shortcoming of ODEs models, that they cannot account for the influence of the microenvironment on the tumour's spatial architecture, its growth dynamics and its response to treatment.

The model developed by Greenspan [236] is one of the first examples of a model that includes a spatial component. It simulates a radially symmetric tumour and a single diffusible growth factor, supplied externally, interpreted as being oxygen. Oxygen concentration within the tumour determines whether local cells die or proliferate. The following equation describes the relationship between tumour volume $V(t)$ and oxygen concentration within the tumour $c(r,t)$ along time:

$$\frac{dV}{dt} = \int_{\partial V} F(c) dV \tag{1.18}$$

where $F(c)$ describes the dependence on local oxygen levels of the net growth rate. Assuming radial symmetry, $V(t) = \frac{4}{3}\pi R^3(t)$, where $R(t)$ is the position of the tumour boundary, equation (1.18) can be reduced to:

$$\frac{dR}{dt} = \int_0^{R(t)} F(c) r^2 dr \tag{1.19}$$

with $R(0) = R_0$. Oxygen distribution follows the following reaction-diffusion equation:

$$\frac{\partial c}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial c}{\partial r}\right) - G(c) \tag{1.20}$$

with these boundary and initial conditions:

$$c(R(t),t) = c_\infty, \ on \ r = R(t) \tag{1.21}$$

$$\frac{\partial c}{\partial r} = 0, \ at \ r = 0 \tag{1.22}$$

$$c(r,0) = c_0(r), \ for \ 0 \leq r \leq R_0 \tag{1.23}$$

where $G(c)$ denotes the oxygen consumption rate, c_∞ the assumed constant oxygen concentration at the boundary, and $c_0(r)$ the oxygen distribution within the tumour at $t = 0$.

Greenspan and others have used simplified forms for $F(c)$ and $G(c)$ such as:

$$F(c) = s_0 c H(c - c_H) - \mu_A - \mu_N H(c_N - c) \tag{1.24}$$

$$G(c) = \Gamma H(c - c_N) \tag{1.25}$$

where H is the function $H(x) = 1$ if $x > 0$, and $H(x) = 0$ otherwise, c_H and c_N are the hypoxic and necrotic thresholds, respectively. Where $c > c_H$ the tumour is well oxigenated and cells proliferate proportionally (factoring in the constant s_0) to the local oxygen concentration, die by apoptosis at rate μ_A and consume oxygen at constant rate Γ. where $c_N < c < c_H$, there is hypoxia, causing cells to not proliferate nor consume oxygen, but to undergo necrosis at rate μ_N. In necrotic regions ($c \leq c_N$, cells are not viable, and they die of apoptosis and necrosis.

Thus, spatial heterogeneity arises from the differentiated oxygenated, hypoxic and necrotic regions. As an example, Lewin *et al.*[237] used this kind of model to emulate the differential response to radiotherapy in well-oxygenated versus hypoxic regions.

Other spatially-determined phenomena can also be modelled using PDEs. For example, Swanson *et al.*[238239] modelled how glioma cells' motility and proliferation rates are different in grey and white matter when invading brain tissue. The model, known as proliferation-invasion, assumes cells move by random motion and can be expressed as:

$$\frac{\partial c}{\partial t} = \nabla(D(x)\nabla c) + \rho c \tag{1.26}$$

where $c(x, t$ is the concentration of tumour cells at time t and position x, ρ is the net proliferation rate of the cells, and $D(x)$ denotes the diffusion coefficient. $D(x) = D_g D_w$ in grey and white matter, $D_w > D_g$, representing the greater motility in white matter. ρ can be replaced with a logistic growth model:

$$\frac{\partial c}{\partial t} = \nabla(D(x)\nabla c) + \rho c\left(1 - \frac{c}{K}\right) \tag{1.27}$$

where K denotes the carrying capacity and ρ the constant growth rate[240].

The proliferation-invasion model has been used to model radiotherapy effects using the LQ model[241], and, with several PDEs extensions, to account for hypoxia, necrosis and angiogenesis[227,240]. It has also been extended to simulate differente treatment combinations[242], and mechanical stimuli, such as in multiphase models. In these models, the tumour is represented as a collection of interacting phases, with their own mass and mechanical properties. They have been used to study vascular tumour growth[243], the response of avascular tumours to different therapies[244,245], or the mechanical influence of surrounding tissue on growth[246].

Phase-field models have been developed as an alternative to multiphase models, its main advantage is that they can include topological changes caused by the shrinkage of a tumour after therapy, or vascular development during angiogenesis[247,248].

1.2.3.4 Agent-based models

In the previous section, I have explained how PDEs can be used to model spatial heterogeneity. However, one of their main shortcomings is that they cannot model tumours at the level of individual cells, which can be key to explain tumour heterogeneity. Agent-based models (ABMs) were developed to account for the effect of subcellular processes and individual cell phenotypes on the tissue-scale processes. Tumours are multiscale, cells influence and are influenced by the microenvironment, the blood vessels, their position within the tumour, and many other components at every level. ABMs seek to represent this with cell-scale spatial resolution. ABMs are often hybrid: they employ a mix of continuum and discrete approaches to simulate the different biological levels. Cells are modelled as discrete entities, their subcellular processes can be stochastic or governed by deterministic ODEs. The distribution of nutrients and drugs can be represented using PDEs, with cells consuming and producing diffusible substances[225]. The spatial position of the cells can be represented as being on a lattice, where they take spaces on a regular grid, such

as in cellular automata or Potts models[249]; or in a lattice-free manner, where cells can be freely placed anywhere, as in Voronoi methods, overlapping-spheres models, and vertex-based models[250]. Cell shapes can be taken into account or ignored, viewing the cell as a single point in space (representing the center of said cell).

As mentioned, lattice models restrict the location of cells to a discrete set of points on a grid. This has several advantages, such as reduced computational cost, allowing the representation of cell heterogeneity while avoiding the complexity of physical and mechanical factors.

The simplest on-lattice models are cellular automatons (CAs). In these, each cell occupies a single point in the grid. This occupancy is updated at discrete timesteps, using a system of probabilities: for example, a cell may proliferate with probability p_1, creating a new cell in an adjacent space, die with probability p_2, by being removed from its place, or it may do nothing with probability $1 - p_1 - p_2$. These probabilities can be specified by cell type, or depend on exterior factors such as nutrient concentration. A CA model was recently used by Brüningk *et al.*[251] to study therapies combining radiation and hyperthermia. The model represents the cell cycle using a timer, but cells can enter a quiescent state if there are no empty adjacent sites, and resume the cycle when these become available.

Potts models are similar to CAs, but in them each cell can occupy a set of several adjacent spaces. As an example, a Pott model was used to study the effect of radiation on glioma stem cell division kinetics[252].

Lattice models are computationally efficient, but they cannot represent complex spatial interactions between neighbouring cells or the effects of mechanical forces, such as intercellular adhesion or volume exclusion. Instead, lattice-free models can be used for this, as in the study by Kempf *et al.*[253], which explored the relationship between radiation and cell cycle synchronisation. For this model, they used a Voronoi-Delaunay method, where cell shape is determined by the position of its centers in relation to the other cells. A subcellular model determines the progression of the cell through the cell cycle and whether the cell passes each checkpoint, pauses cycling or dies by apoptosis.

Another recent example by Wang *et al.*[254], uses an off-lattice representation to build a quantitative systems pharmacology model built on published data of TNBC. They generated a virtual patient cohort to conduct an *in silico* clinical trial, using the results of a real clinical trial for calibration and validation. The calibrated model was used to predict efficacy and biomarkers of response.

As ABMs become more sophisticated, so their computational cost and development time increase. Because of this, a number of freely available modelling frameworks have been developed, platforms on which specific models can be developed with the necessary features and parameters, with the advantage of having all the common features, such as substance diffusion or spatial placement already implemented in a robust manner. PhysiCell[255] is one of such frameworks, it employs a lattice-free, physics-based approach and a range of cell behaviours such as cycling, death, volume regulation, motility and mechanical interactions between cells, and combines it with BioFVM[256] to include interactions between cells and diffusible substances such as oxygen. Another example is microC[257], a framework that combines a lattice cell model, where every cell is represented by its center point, a diffusion system, and a Boolean Regulatory Network (BRN) to account for subcellular processes. Such networks are graph representations of gene and protein interactions, with every element represented as a node with a Boolean function that integrates the incoming edges and determines the state of the node (active or inactive). Output nodes such as proliferation or apoptosis serve as a readout of the cell state, and effect changes at the cellular level. The modelling framework I have developed as part of this thesis, *microCJ*, is inspired by microC, but its implementation is more robust, its design is modular, so other scientists can more easily learn, use and extend it, and it adds key features, such as lattice-free placement, drug treatments, or the modelling of subcellular processes as Probabilistic Boolean Networks (PBNs), which allows for continuous representation of protein levels. The design, implementation and uses of *microCJ* will be thoroughly discussed in chapter 3 and chapter 4.

Other examples of ABM frameworks and tools include Cancer, Heart and Soft Tissue Environment (Chaste)[258] wich implements a range of different ABMs,

Hybrid Automata Library (HAL)[259], with modular hybrid modelling features and visualisation tools, Morpheus[260] and Biocellion[261].

1.2.3.5 Other examples of studies using mathematical modelling

In glioblastoma, an aggressive brain tumour, the vasculature is disorganised and presents a disrupted blood–brain barrier (BBB). It is not clear if the restoration of BBB integrity resulting from antiangiogenic therapy improves or limits the delivery of drugs. Boujelben *et al.*[262] applied mathematical models of blood flow, vascular permeability and diffusion within the tumour microenvironment, and found that these three parameters interact non-linearly to determine the concentration of drug in any given point. Restoration of the BBB was predicted to reduce the delivery of the drug, but higher blood flow could increase it.

Karolak *et al.*[263] developed an off-lattice agent-based model to examine which biochemical and biophysical properties are essential for effective spatial distribution within the tumour and cellular uptake. They predicted that drug agents with moderate affinity, but that are quickly released, are similarly effective to high affinity drugs released slower.

Targeted therapy has also exploited nanoparticles (NPs) technology. Curtis *et al.*[264] studied an important factor in its effectivity, namely the NP vascular affinity, which represents the likelihood of endothelial adhesion. They used a hybrid model with a discrete vasculature, continuous description of tumour mass and drug kinetics. Their results predicted that small (100 nm) NPs with high vascular affinity should be more effective than larger (1000 nm) NPs with similar affinity, even though small NPs have lower drug loading and local drug release.

Macrophages accumulate in hypoxic regions of tumours. This makes them promising candidates for targeting cancer cells in those areas, where conventional drugs might not be able to reach, and where their efficacy is anyway limited in many cases. Owen *et al.*[265] developed a cellular automaton model that incorporates spatial and temporal dynamical parameters of vascular tumour growth. This model showed that the combination of conventional therapies with macrophage therapies

was predicted to be synergistic, and that the greatest effect would be obtained where the macrophage-based, hypoxia-targeted therapy is administered shortly before or concurrently with conventional chemotherapy.

Another modelling effort in this area has been published by Leonard *et al.*[266], who studied how macrophages can be used to deliver nanodrugs to hypovascularized metastatic lesions. Their model combines the simulation of the evolving tumour, a representation of the vasculature and diffusion equations for the drug distribution. Both the model and the *in vitro* and *in vivo* experiments showed that the addition of macrophages significantly increases the drug efficacy.

Lastly, a very promising area in cancer treatment which can also be studied through mathematical modelling is combination therapy. For example, there is substantial evidence that chemotherapy and radiotherapy, when combined, can achieve a greater effect than that of any of these therapies alone[267–269]. However, the success of this kind of interaction is greatly dependent on the exact parameters of the combination: namely the dose of both chemotherapy and radiotherapy, the number and frequency of radiation fractions, the chemotherapy schedule, and the type of radiation and drug administered[270]. The large number of possible combinations is the reason why mathematical modelling can be of great help to evaluate different multimodality strategies before testing on patients. Several efforts have been made in this respect[270,271], but models that explore in detail the interplay between the effects of combination therapy, as well as the optimal scheduling and sequencing of the combination treatment are still a work in progress[272].

1.2.4 Opening the cell's "black box"

Cells can be thought of as entities where that take in inputs and return outputs in response. Biological models have tended to treat the cell as a "black box"[273]: they often ignore the interplay between the environment around the cell and the events inside it which determine the specific output. The reality is that cells are vast networks with thousands of components interacting, complex economies in which scarce resources are carefully invested depending on fluctuating environmental

and cell intrinsic conditions[274]. In addition, gene mutations can have subtle but unique effects on the whole network, leading to diverse phenotypes across cancer patient populations[275]. At the same time, cells do not exist in a vacuum, but they coevolve with their environment[276,277]. Modelling gene networks in isolation ignores the impact of the cell context on its phenotype. It has been argued that establishing a direct relationship between single genes and phenotypic traits is an oversimplification, since in most cases no sole gene can cause any observable difference[277]. And of course, genes are not the sole influence on phenotypes, they depend on a cellular environment and on specific physicochemical conditions.

Cancer genomes are very complex, including tens to hundreds or more somatic alterations in each tumour[275]. Although in the past genetic mutations in cancer have been approached as if they entirely change gene function, it must be noted that each mutation could influence protein function or affect networks in subtle ways, and different mutations in the same gene can lead to different phenotypes[278]. All the above contribute to increase diversity across cancer patient populations even further.

microCJ, my modelling framework, tries to address these needs, and merge pathway to spatial modelling. In it, models combine the 3D architecture and conditions of the tumour microenvironment, the genotype inside each of the cells, and the signalling networks determining the responses to environmental stimuli. In this view, collective behaviour and phenotypes rise from the profiles of individual cells, which rely on the simulation of metabolic and genetic networks inside each cell. These individual behaviours can mutate and evolve, allowing one to study causal relationships between the system emergent properties and the genotype of each cell.

1.2.5 Data explosion and open access repositories

High-throughput technologies and genomewide analyses have provided us with enormous amounts of biological data that are measured in petabytes, with estimates indicating that by 2025, one will need more than two exabytes of storage capacity just for human genomes[279]. Some of these resources are dedicated to cancer, like The Cancer Genome Atlas (TCGA)[280] and the International Cancer Genomics

Consortium (ICGC)[281], and it is clear that cancer models can benefit immensely from using them. Some projects, like TCGA and ICGC, aim at providing large-scale multi-'omics characterization of cancer samples from individuals and integrate this with demographics and clinical information (e.g. disease stage, clinical outcome), while others like the Library of Integrated Network-based Cellular Signatures[282], the Integrated Cancer Biology Program[283], or the Cancer Dependency Map[284] have characterised the molecular response of several cancer lines to a range of different perturbations. Furthermore, many other resources exist which are not necessarily linked with cancer, such as repositories of biological pathways (KEGG[285], WikiPathways[286], Pathway Commons[287]), but could also be used to inform modelling of signalling network in normal and cancer cells.

This abundance of data comes with many perks, of course, but it is not devoid of challenges. A recurring theme in biological data science is that currently the ability to integrate and interpret data across platforms lags behind the ability to generate data[288]. First, one needs to sieve through the data, determine the level of evidence or the noise associated with the deposited data, and recognise what could be relevant to usefully fit cancer models. Secondly, none of the platforms is comprehensive enough to capture the whole array of changes in the genome of a cancer cell, or to properly differentiate between driver and passenger mutations[289]. Thus, these collections present different formats and paradigms, and it is quite a challenge to integrate several sources of information and to transform them to formats that can be fitted to the model, since they are not always directly comparable. To attempt to address these problems, several computational tools and databases have been developed[289,290].

There are some new ways to approach the analysis of such vast amount of data, like crowdsourcing methods, which have already produced promising results[291,292]. Despite the problems that inevitably arise from large-scale collections of data, making these available to the community makes it easier to detect errors. Sharing of data and crowd-modelling also prevents the self-assessment trap, by which models are built and tested by the same group, with the risk of partiality in the results[289].

1.3 Aims of this thesis

In light of the unmet needs exposed in this literature review, the aims of these thesis are:

- To develop a multi-scale computational model of TNBC that integrates current knowledge of gene regulation and biochemical pathways.
- To use transcriptomic data from TNBC patients and cell lines obtained from *in vitro* and *in vivo* samples to inform the computational model.
- To use this novel computational model to gain insights into resistance to EGFR-targeted therapy and to point to future strategies for combination therapy.

2

Analysis of single-cell and spatial transcriptomics

As it was described in section 1.1.1, TNBC is a cancer type of unmet needs. EGFR is an intriguing target for this type of cancer due to its resistance to treatment in TNBC, despite its success in other tumour types (see section 1.1.2.3). Several published transcriptomic studies are accessible that relate to this specific cancer subtype, which are attractive experimental test cases to apply my computational model, *microCJ*, in the study of the pathways involved in resistance to EGFR-targeted therapy in TNBC. The results of the simulations with *microCJ* after fitting it with transcriptomics data are the subject of the last chapter of this thesis. But first, I will dedicate this chapter to the analysis I performed of these datasets.

2.1 An EGFR-dependent tumour-initiating program in TNBC

The authors of Savage *et al.*[1] kindly allowed me access to the data described in their paper. In it, after screening TNBC patient-derived xenografts (PDXs), they identified an exceptional responder to treatment with gefitinib, the EGFR inhibitor. Single-cell analysis of cells from this PDX showed several subpopulations with distinct features. A mesenchymal/stem-like cluster displayed high expression

of EGFR and tumourigenic potential. There is increasing evidence that supports the idea that tumour-initiating cells (TIC) reside at the apex of a tumourigenic hierarchy where they both self-renew and give rise to a proliferative population that differentiate toward cell with diminished tumourigenic capacity,[293] and therefore Savage *et al.*[1] hypothesise that in the exceptional responder, these tumour-initiating cells overexpress EGFR ($EGFR^{hi}$ cells) and give rise to non-TIC cells with reduced EGFR expression ($EGFR^{lo}$ cells).

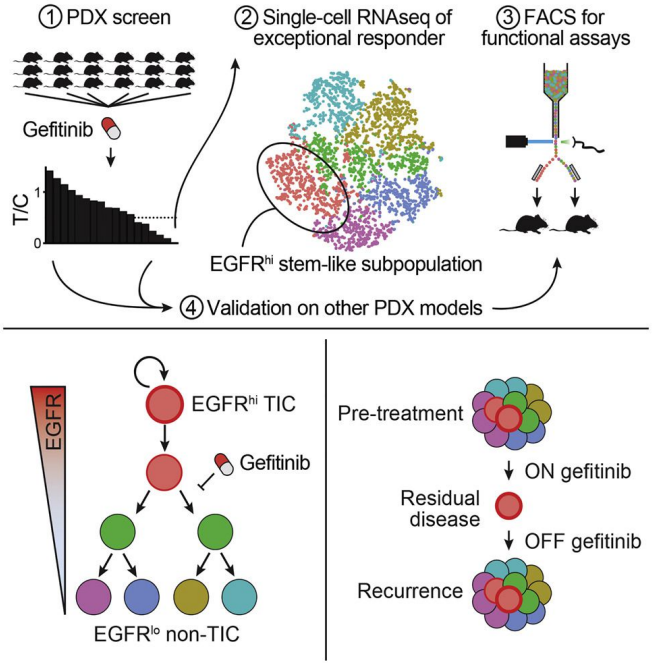


Figure 2.1: A graphic abstract of the study carried out by Savage *et al.*[1]. *Top:* The experimental design of the project. *Bottom:* The EGFR heterogeneity-dependent hierarchy found in the exceptional responder, where $EGFR^{hi}$ cells are at the top, self-renewing and producing other subpopulations of $EGFR^{lo}$ cells, this transition is blocked by gefitinib. Figure by the Savage *et al.*[1] authors.

Moreover, the authors found similar tumourigenic $EGFR^{hi}$ subpopulations in independent PDXs, where EGFR heterogeneity correlated with gefitinib sensitivity. The paper hypothesises that there exists an EGFR-dependent program where EGFR heterogeneity and tumour initiating capacity are associated and it is sensitive to targeted treatment, and this is supported by several experiments the authors performed. The proliferative transition from $EGFR^{hi}$ to $EGFR^{lo}$ is inhibited by gefitinib, and although the treatment destroys the $EGFR^{lo}$ subpopulation, a $EGFR^{hi}$

residue remains, which is capable of regenerating the tumour after treatment has stopped, giving again rise to EGFR heterogeneity (figure 2.1).

In the article, the authors only analysed the control scRNA-seq dataset, but they then allowed us access to three single-cell RNA-seq datasets: control — PDX from untreated mice, treatment — mice treated for four weeks, and off-treatment — mice treated for four weeks and then untreated for another four weeks. For a full description of the treatment protocol and dosage, see Savage *et al.* [1].

2.1.1 Analysis

The data was first checked for general quality and the Unique Molecular Identifier (UMI) counts and the percentage of mitochondrial genes were filtered, to a minimum of 10000 counts for the former and 25% for the latter (figure 2.2).

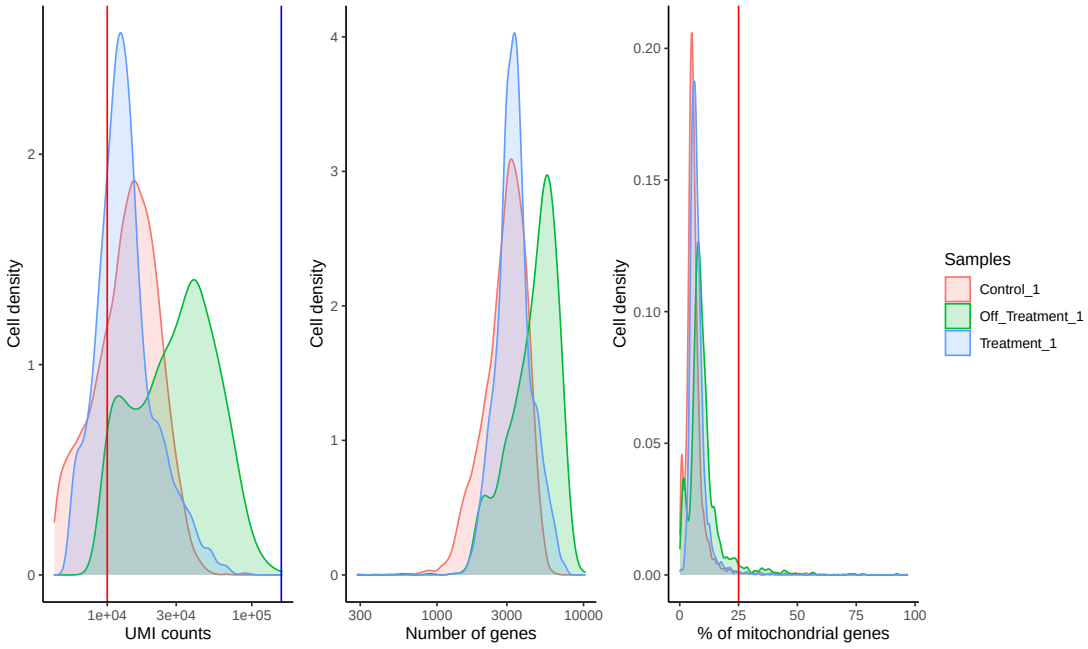


Figure 2.2: Quality control measurements. Unique Molecular Identifier (UMI) counts, number of genes and percentage of mitochondrial genes are displayed for every sample (control, off-treatment and treatment).

I then clustered the cells in 13 subpopulations based on their transcriptomic identity (which are represented in their Uniform Manifold Approximation and Projection - UMAP - in figure 2.3, see Methods). After studying the characteristics of each

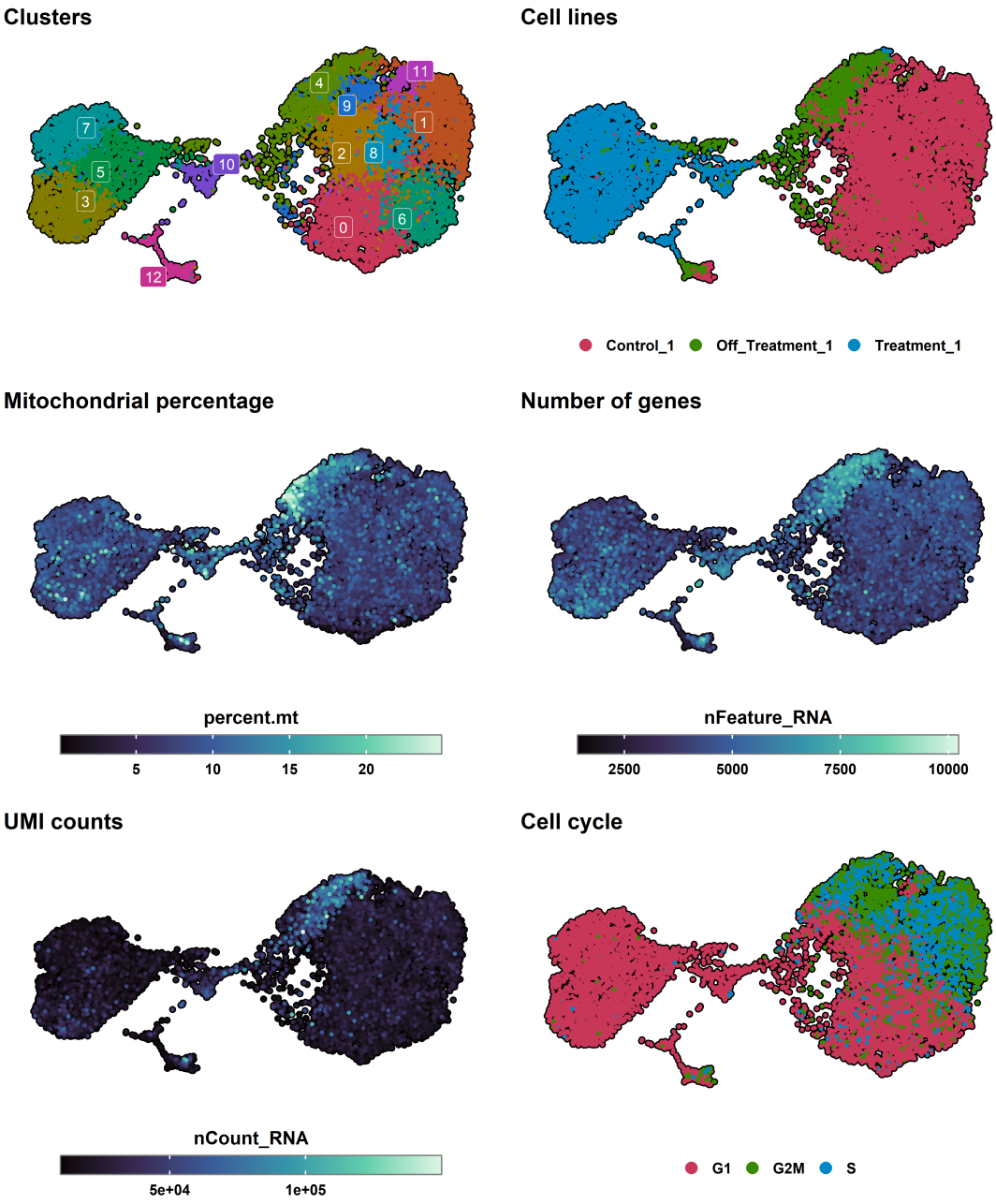


Figure 2.3: Overview of the datasets. The figures show an Uniform Manifold Approximation and Projection (UMAP) representation of the three merged datasets (control, treatment and off-treatment), with cell coloured by different features: cluster number (that will be annotated later), cell lines or samples (control, treatment and off-treatment), percentage of mitochondrial genes ("percent.mt", with quantitative color scale), number of genes ("nFeature_RNA", with quantitative color scale), unique molecular identifier counts ("nCount_RNA", with quantitative color scale), and cell cycle stage (G1,G2M and S).

cluster using various transcriptomic characteristics described in detail in the following paragraphs, I assigned an identity to each cluster. The names are prefixed by C (control, corresponding to seven clusters) T (treatment; four clusters in total) or OT (off-treatment, with one cluster). Clusters are named according to their biological characteristics and TNBC subtype: C-B (basal) 1-3 , corresponding to clusters number zero, two and eight, respectively, from figure 2.3; C-Prolif (proliferation) 1-2, corresponding to clusters one and nine; C-MSL-EGFRhi (mesenchymal/stem-like and high EGFR), corresponding to cluster six; C-IM (immunomodulatory, cluster eleven); T-M (mesenchymal, cluster three); T-B (cluster five); T-MSL-EGFRhi (cluster seven); T-LAR (luminal androgen receptor, cluster ten); OT-Prolif (cluster four), and Transition-MSL (cluster twelve), which is formed by a mix of cells from the three datasets, and present a mixed/MSL expression.

The optimal resolution for clustering was chosen to achieve the minimum possible number of biologically relevant clusters and TNBC subtypes (named as described below according to their gene signatures and pathways) while separating cell lines and cell cycles as much as possible. One can also see in figure 2.4 that at the resolution of 0.6 cells start to cross over from far away branches of the tree, indicating that a higher resolution would mean overclustering the data.

Of note, the clusters with cells predominantly coming from the treatment dataset (seven, five and three at resolution 0.6) stayed stable with different resolution values, while the rest multiply as the clustering becomes more focused (figure 2.4). This shows that the treatment dataset is much more homogeneous than the control (as can also be inferred from the fewer number of cells from the treatment sample that find themselves assigned to cluster identities where other samples are more abundant, see figure 2.5). This homogeneity would be caused by the fact that most cell subtypes are killed by the treatment, and only a fraction of them are present in the treatment dataset. The general transcriptomic landscapes from the off-treatment and control samples are similar, as showed by principal component analysis of the variable genes in figure 2.6 a , where the control and off-treatment mostly overlap

across the main axis of variation, the first principal component (PC). One cluster-sample combination, Transition-MSL from the off-treatment sample, separates from the rest across the second principal component (figure 2.6 b), which nevertheless explains a small percentage of the total transcriptomic variation (three percent).

To aid cluster identification, cells were scored by their cell-cycle phase signature, which tries to classify cells by their likely level of progression through the cell cycle. In figures 2.3 and 2.7 one can see how most cells are deemed to be in the G1 phase, and only 37% of them are in S or G2M. Virtually all cells in G2M or S stage are from the control or off-treatment datasets (figure 2.7), while the treatment clusters do not present such cells at all. The higher fraction of cells in G2M/S stages in the control and off-treatment samples suggests that the cells might be more actively dividing (proliferating) than in the treatment stage, although the genes used for scoring the G1 stage cannot actually distinguish between quiescent cells (G0) and those proliferating cells preparing for cell division.

Then I looked at the EGFR levels of expression. Figure 2.8 shows that $EGFR^{hi}$ cells those thasss present EGFR counts higher than 2 concentrate mostly in cluster C-MSL-EGFRhi for the control and off-treatment datasets, and in cluster T-MSL-EGFRhi for the treatment one (figure 2.8 c). This would support Savage *et al.*[1] description of an $EGFR^{hi}$ population (C-MSL-EGFRhi) that remains alive during treatment (T-MSL-EGFRhi), although possibly with inhibited growth, as shown by the lack of G2M/S cells and as hypothesised by Savage *et al.*[1] (figure 2.7).

Still, the hypothesis put forward by Savage *et al.*[1] is not the only possible explanation, since it would also be possible that the same mutation is being selected at different points in the tumour history. This provides a good example of why dynamic, multiscale computational models are useful: they can provide mechanistic explanations to help us disambiguate what is the most likely hypothesis, and narrow down the necessary experiments to confirm these explanations. In an attempt to further define the clusters I found, I identified the most significant differentially expressed genes of each cluster, calculated against the gene expression average of all others(figure 2.9), and performed pathway enrichment analysis (figure 2.10) using

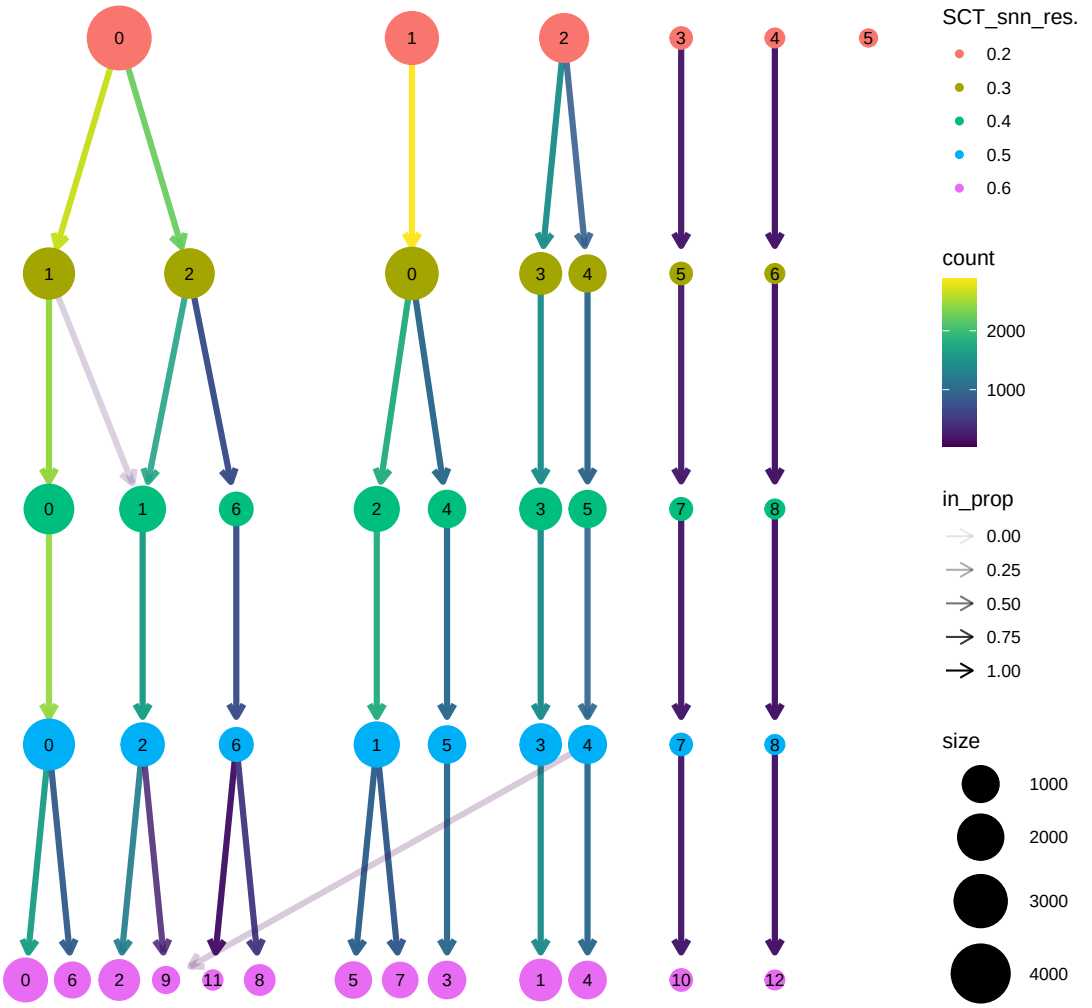


Figure 2.4: Clustering tree of Savage *et al.*[1] data. the cluster tree shows how the data is clustered using different resolution parameters. The number of clusters increases with the resolution: stable clusters are maintained with increasing resolution (cells do not split or cross over to far away clusters in the tree). The resolution itself is represented by the changing colors in the descending rows (legend "SCT_snn_res"). How cells are reassigned from one cluster to another as resolution increases is indicated by the arrows. The proportion of cells moving to the new cluster is denoted by the width of the arrow (represented by the transparency of the arrow in legend labelled "in_prop"). The number of cells in each cluster is proportional to the circle area.

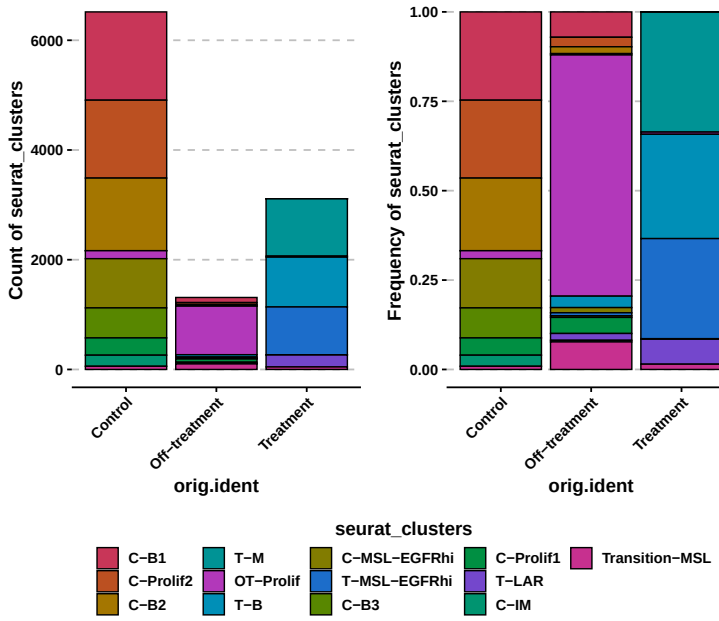


Figure 2.5: Cell abundance per cluster in Savage *et al.*[1] data. Each bar represents the cells in one of the three datasets (control, treatment and off-treatment) and their cluster composition (colors). The left plot represents the total number of cells in each cluster and the right one plots the clusters by their size relative to the total. This figure shows that the treatment and off-treatment samples contain half the cells or less than the control, and the treatment sample is more homogeneous (contains cells from a smaller variety of clusters than the control and off-treatment).

Progeny[294]. This gives each cell a score of how over-represented a cancer pathway signature gene set is among the expressed genes, compared to a random set of the same size. Overall, this analysis shows that most proliferation-related pathways (MAPK, PI3K, VEGF), even the EGFR downstream, are repressed in the treatment dataset and active in the others. Again, this reinforces the theory that treatment inhibits the growth of the tumour, but it does not kill the $EGFR^{hi}$ cells.

Cells were also classified by their transcription factor score (see Methods), shown in figure 2.11. Again, this shows the proliferation activation-inhibition difference between treatment and non-treatment. TFs related to proliferation such as MYC are repressed in treatment but active in the other clusters. On the other hand, tumour suppressor TFs, like EGR1 or FOXO3, are active in treatment and inactive otherwise.

Finally, cells were scored using Lehmann *et al.*[21] subtypes signature[295] (figures 2.12

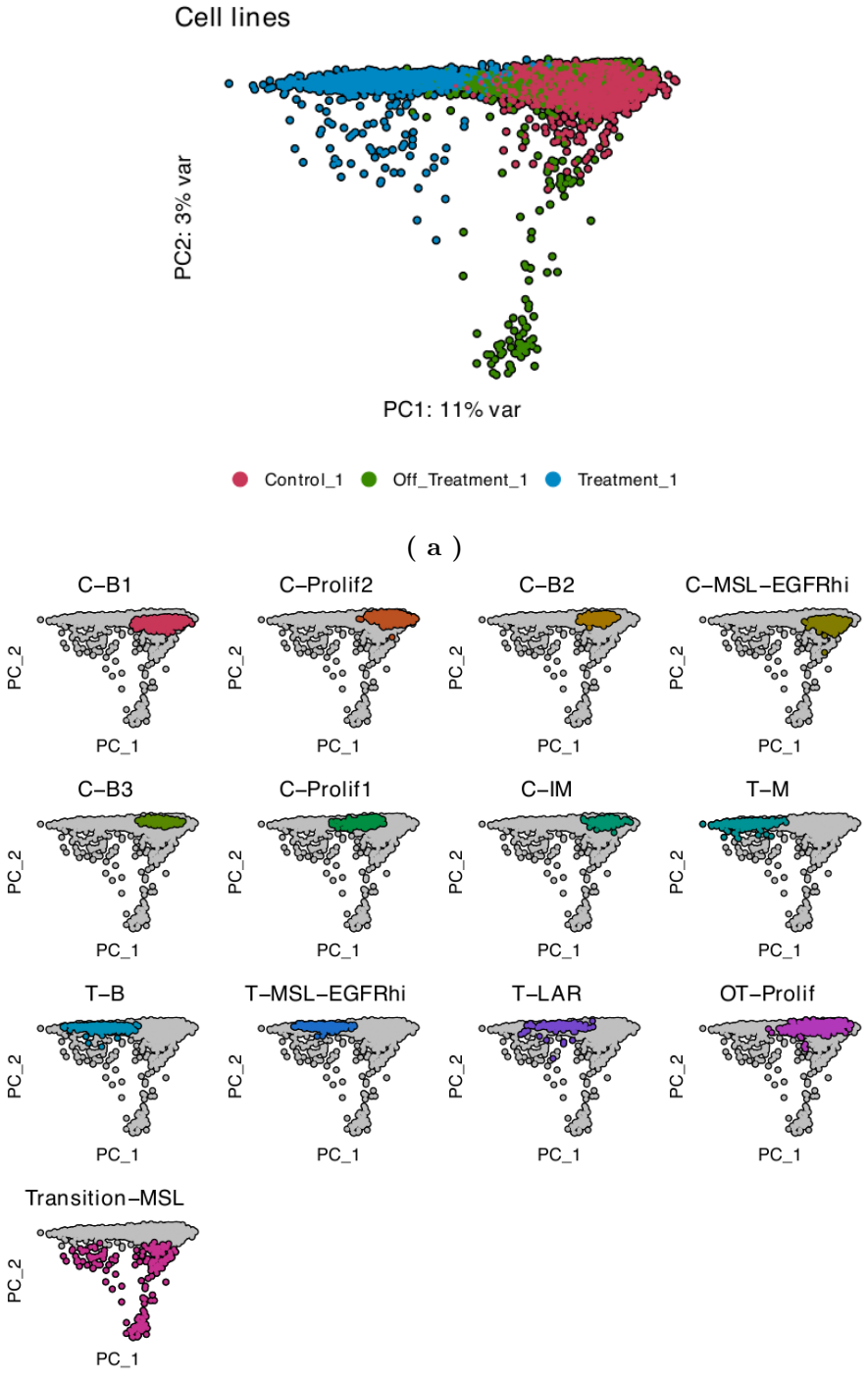


Figure 2.6: Principal component analysis. (a): Distribution of the cells, colored by sample, across the first two principal components (PC) of the transcriptome, including the variance explained by each PC. (b): Same as in a), but colored and split by cluster.

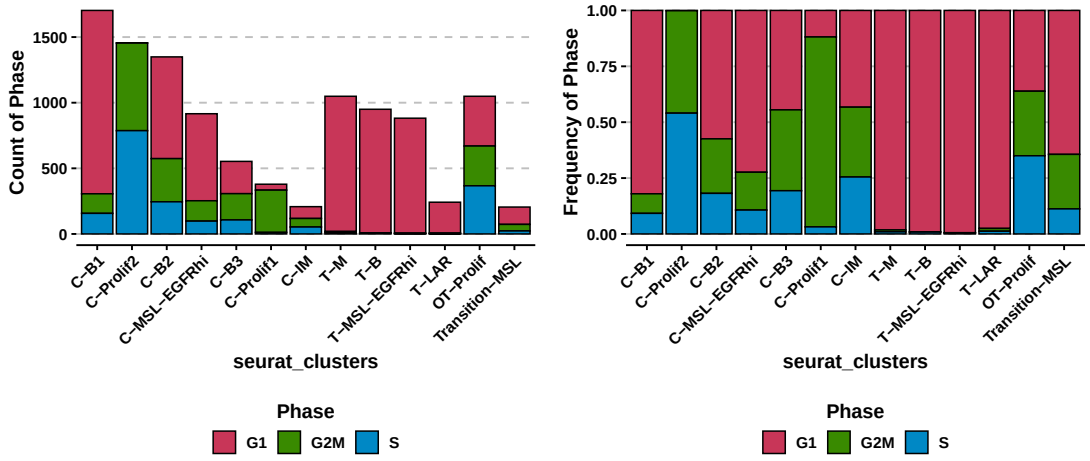
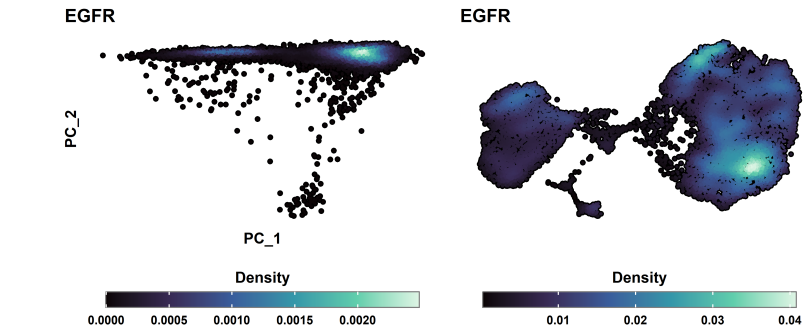


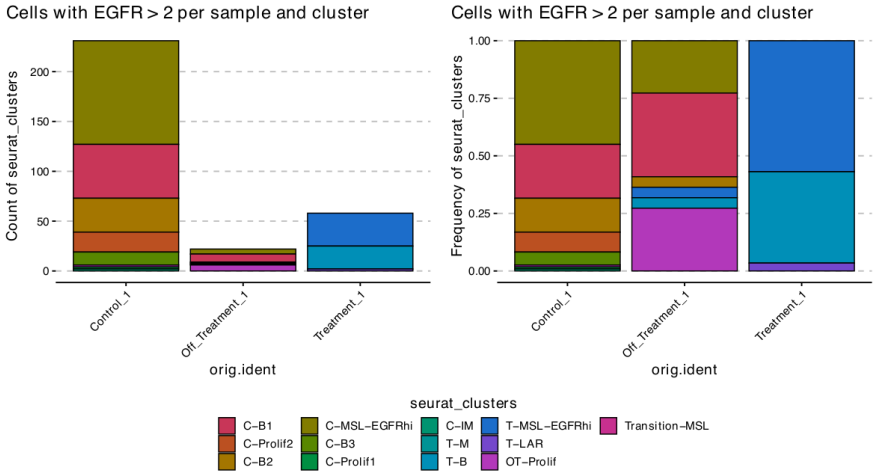
Figure 2.7: Cell abundance per cell cycle phase in Savage *et al.*[1] data. Each bar represents the cells each one of the clusters, and their cell clycle phase composition (colors). The left plot represents the total number of cells in each cluster and the right one plots the clusters by their size relative to the total. This figure shows that the clusters from the control (C) and off-treatment (OT) samples contain the vast majority of cells in G2M and S cell cycle stages, compared to the treatment (T).

b and 2.13, and the one Savage *et al.*[1] adopted in their article (figure 2.12 a). This scoring was used to ascertain the identity and biological meaning of each cluster. The relative enrichment of each cluster would further support Savage *et al.*[1] description of the $EGFR^{hi}$ populations as mesenchymal/stem-like, since after basal-like 1 and 2, which almost all clusters score highly for, is the highest scored subtype for this cluster.

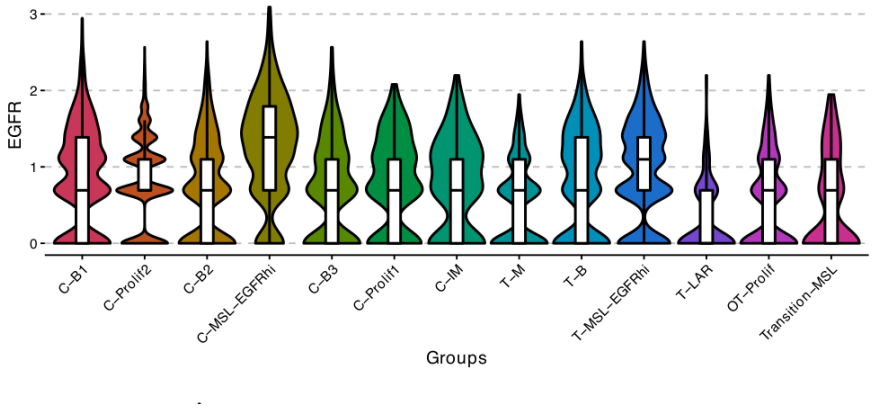
Taking all this information together, I annotated the clusters. The initial cluster zero presents a Basal signature, as shown by figure 2.12 a and figure 2.12 b , and thus was named C-B1. The same reasoning was used to annotate cluster five as T-B.. Cluster one presents a proliferation signature from figure 2.12 a , and a majority of cells in phase G2M and S, and thus was classified as proliferative, along with cluster nine (C-Prolif 1 and 2, respectively). Cluster three presented a mesenchymal profile as shown by figure 2.12 b (thus named T-M, although it also had a high Basal signature by Savage's classification). Cluster six and seven show mesenchymal/stem-like signatures and high EGFR expressions in control and treatment samples (C-MSL-EGFRhi,T-MSL-EGFRhi), respectively, as can



(a)



(b)



(c)

Figure 2.8: EGFR status. (a): PCA and UMAP representations with cells coloured by their relative EGFR expression. The plots are coloured by the density of cells expressing EGFR, which takes into account whether the neighbour cells are also expressing it. (b): Number of cells with relative EGFR expression over two depth-corrected counts, grouped by dataset and colored by cluster. The left plot represents the total number of cells, and the right one the frequency. (c): Expression of EGFR by cluster and dataset (C: control, T: treatment, OT: Off-treatment) plotted as violin plots with a box plot inside.

be seen in figure 2.8 c . Cluster twelve was classified as mesenchymal/stem-like (T-MSL) from figure 2.12 b . Cluster ten was classified as luminal-AR (T-LAR), and cluster eleven as immunomodulatory (C-IM), also by Lehmann's signature. Cluster four was classified as proliferative (OT-Prolif) attending to Savage's signature in figure 2.12 a (although it was difficult to classify due to the weak enrichment) and to the large percentage of G2M/S cells. Finally clusters two and eight were classified as basal (C-B 2 and 3) attending to Lehmann's classification, although the basal enrichment was weak using Savage's classification.

These results are relevant to the rest of the project because it provides a platform to analyse experimental data, taken at three different time points, using computational modelling with *microCJ*. Further analyses using *microCJ* starting from the control RNA-seq dataset can be then compared to the other two datasets and see if the *in silico* tumour and the real one evolve in the same way.

Additionally, whereas Savage *et al.*[1] only analysed the control dataset, I have also analysed the treatment and the off-treatment one, and an interesting result of this analysis is that these do not contradict the authors's hypothesis of the resistant EGFRhi population, although more experiments would need to be conducted to fully corroborate it.

I then carried forward the cell-level gene expression matrix from each cluster as experimental input for *microCJ*.

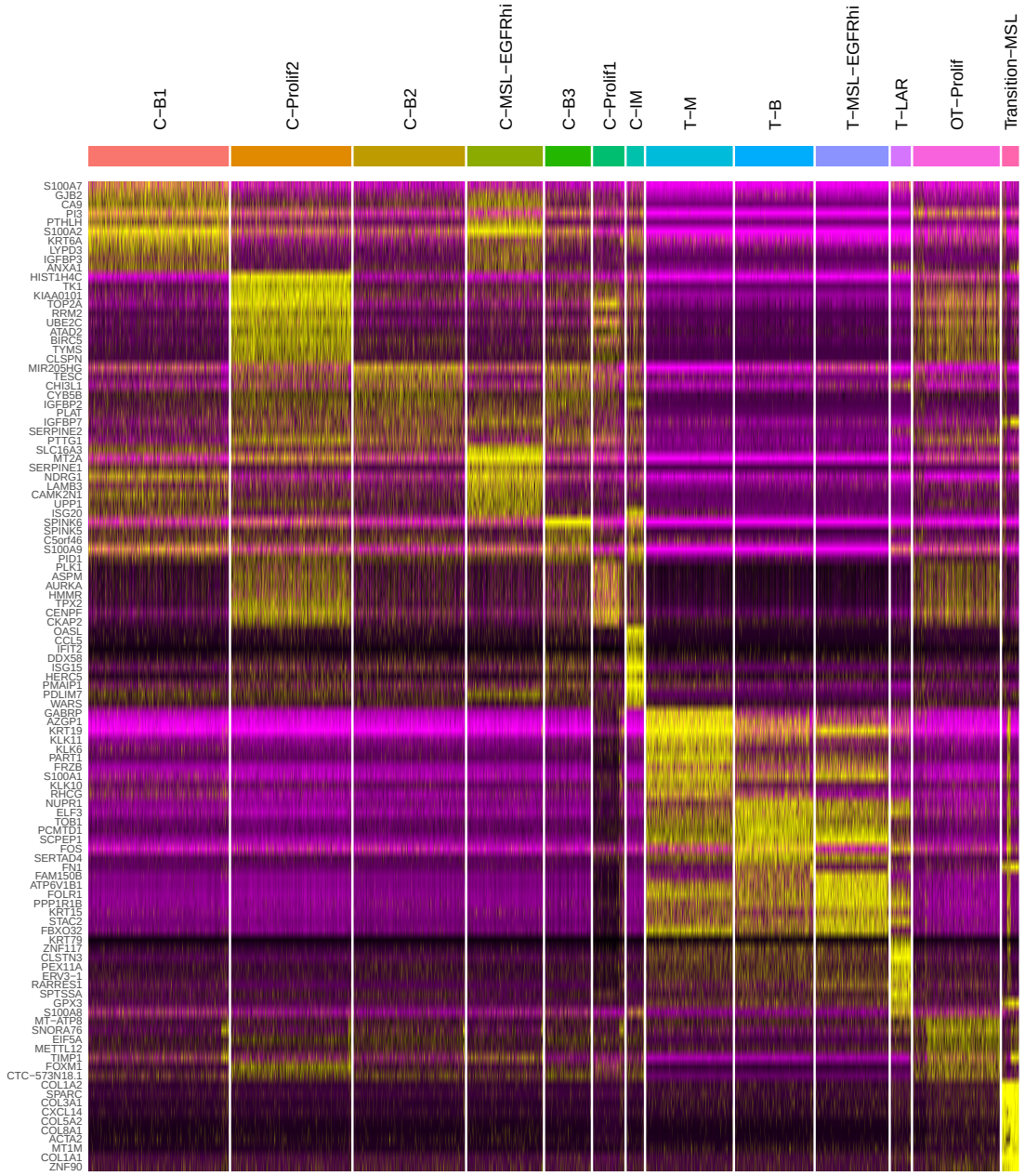
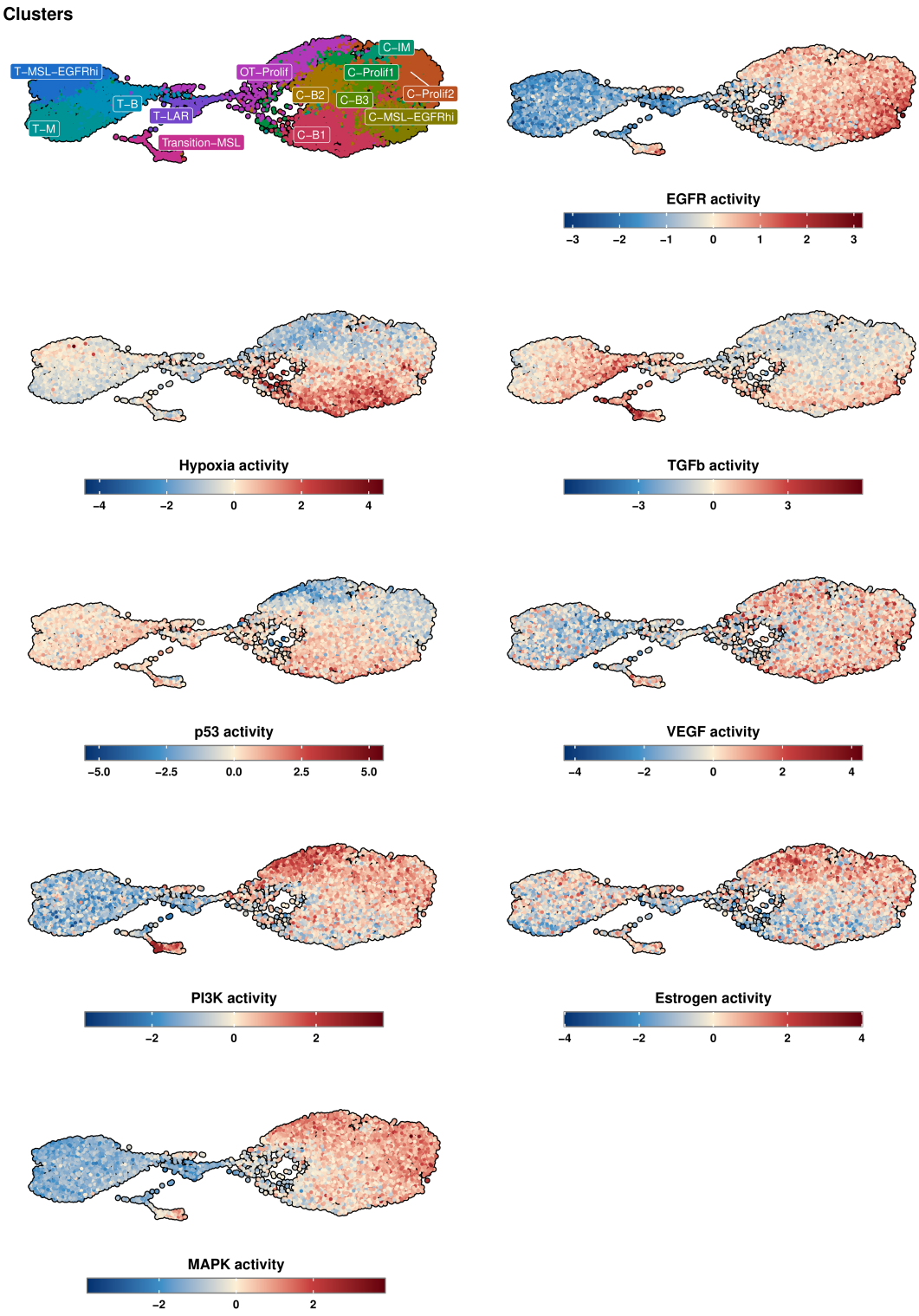
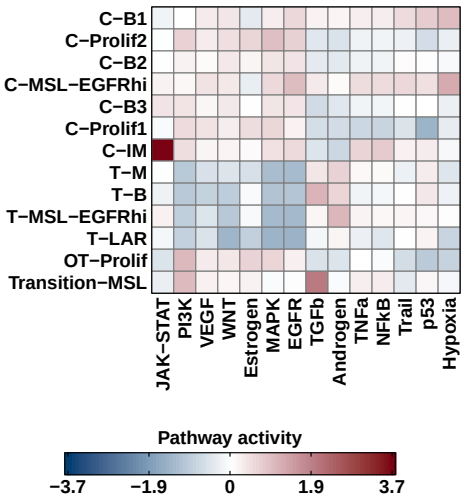


Figure 2.9: Marker genes. Heatmap showing the top 10 marker genes per cluster. Markers were calculated by differential expression analysis of each cluster against all others, genes with positive log2 fold change per cluster were kept and sorted by increasing Bonferroni-adjusted p-value. Yellow colours indicate an increased expression, and purple a decreased expression.


Clusters
T−MSL−EGFRhi
T−B
T−LAR
T−M
Transition−MSL
OT−Prolif
C−B2
C−B1
C−IM
C−Prolif1
C−B3
C−Prolif2
C−MSL−EGFRhi
EGFR activity
−3 −2 −1 0 1 2 3
Hypoxia activity
−4 −2 0 2 4
TGFb activity
−3 0 3
p53 activity
−5.0 −2.5 0.0 2.5 5.0
VEGF activity
−4 −2 0 2 4
PI3K activity
−2 0 2
Estrogen activity
−4 −2 0 2 4
MAPK activity
−2 0 2

(a)



(b)

Figure 2.10: Cancer pathways enrichment. (a): UMAP representation with cells coloured by several pathways activity. (b): heatmap showing the pathway activity score by cluster.

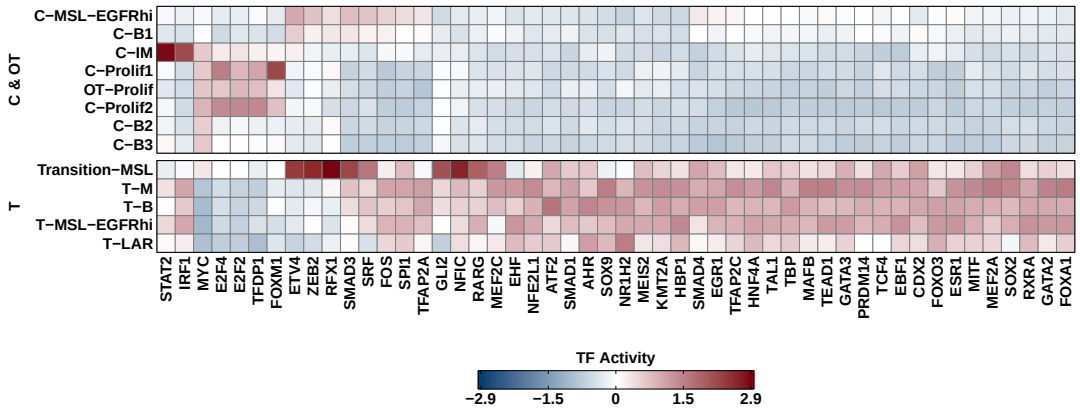


Figure 2.11: Transcription factor activity score. Heatmap showing the transcription factor activity score by cluster. Cluster are grouped by the dataset their cells mostly come from, treatment (T) or control and off-treatment (C&OT).

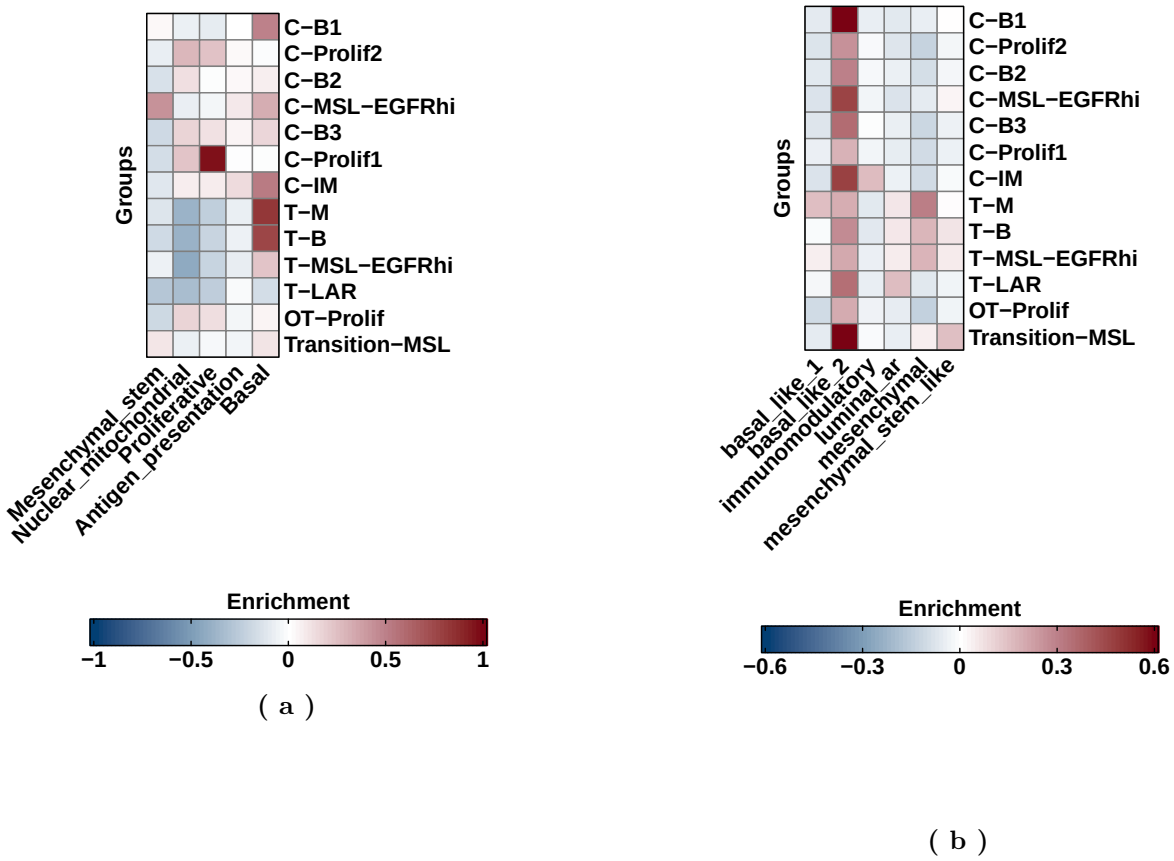


Figure 2.12: Heatmaps showing the enrichment per cluster for two different signatures. (a): signature described in Savage *et al.*[1]. (b): signature as per Lehmann *et al.*[21]

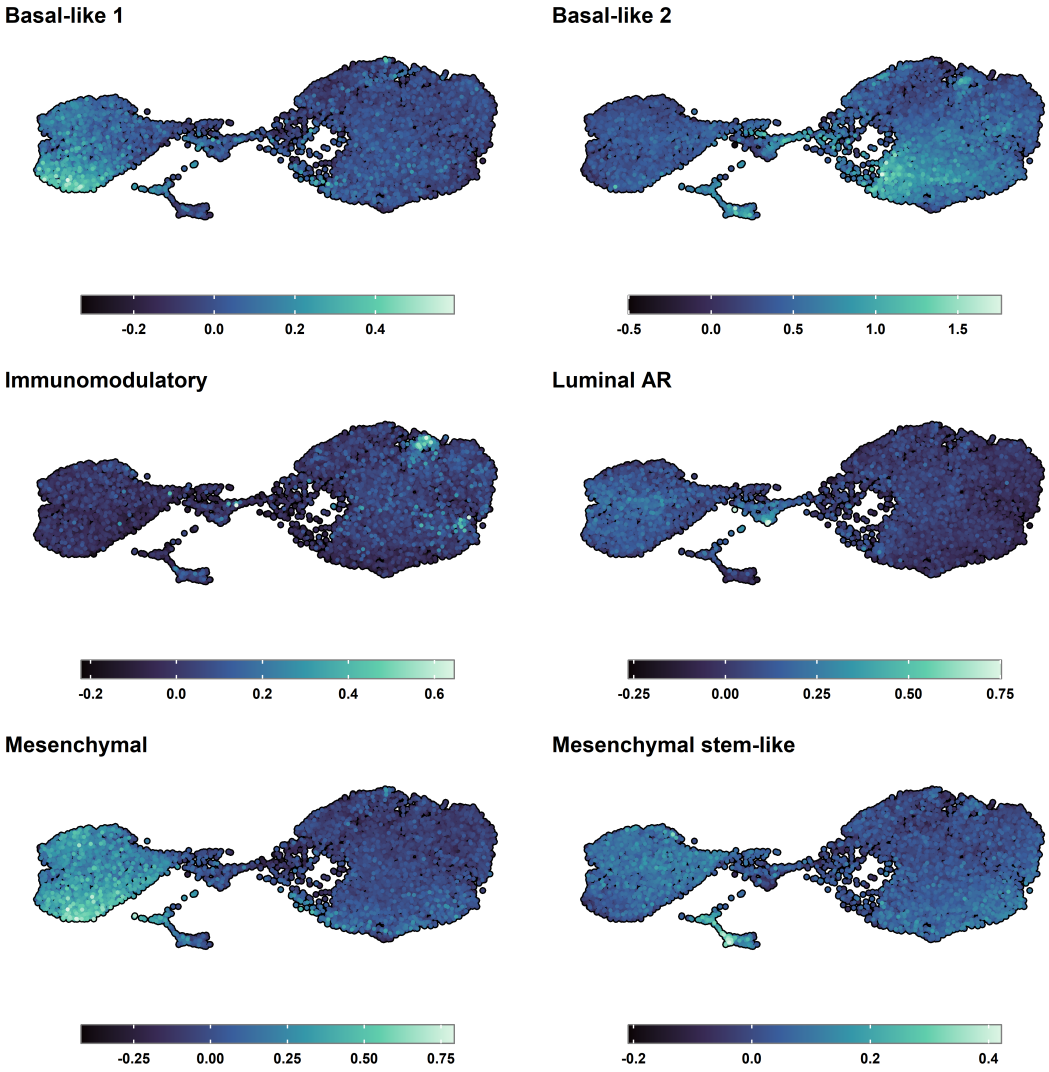


Figure 2.13: TNBC subtype scoring. UMAP representation wich cells coloured by their TNBC subtype score, attending to Lehmann *et al.* [21] classification.

2.2 Spatial transcriptomics

I was allowed access to spatial transcriptomic data from four tumor biopsy from untreated samples from ER negative, PR negative and HER2 negative breast cancer. Fresh cancer biopsies were flash frozen and subsequently processed between 2020 and 2021 by several collaborators and colleagues (see Methods), coordinated by my colleague Dr Karen Sayal. These samples were collected from three participants (named BRO18, BRO20 and BRO23) of the BRECO study[296], as four 4 mm tumour punch biopsies (referred to as "pucks") and sequenced with the spatial transcriptomics Slide-seq technique.

The initial bioinformatics analysis was performed by Clara Pavillet. She performed the initial quality control of mitochondrial percentage and UMI counts, and cells were filtered to those with over 1000 UMI counts and below 25% of mitochondrial genes. The pre-filtered distributions are shown on figure 2.14, with thresholds highlighted. After filtering, pucks one and four contained 41 and 35% percent of the original cells, 6,423 and 4,868 cells respectively, compared with pucks two and three which retained 15,213 and 20,925 cells respectively. Normalisation, scaling and dimensionality reduction was then performed (see Methods).

I then clustered the cells and built cluster trees (figure 2.15). The final resolution chosen for clustering of all samples was 0.1, as a smaller resolution meant that there were too few clusters in pucks one to three to separate biologically meaningful subgroups. On the other hand, at a 0.1 resolution cells start crossing over from clusters further away in the tree in pucks one, two and four, which means that overclustering is starting to occur, which including the fact that some clusters could not be annotated (as is explained below) based on their genetic signatures justified not using a larger resolution. This resulted in eleven clusters for puck one, eight clusters for pucks two and three each, and fourteen clusters for puck four.

One can also perceive from the cluster tree that puck four from sample BRO23 presents a higher genetic heterogeneity, since there are more clusters with a more distinct identity (signalled by the vertical and separate branches of the cluster tree). The higher amount of cells crossing over in puck one and three, coupled with the

lower number of cells retained after count quality control, led to these two pucks and samples being dropped from further analysis.

I then performed the signature enrichment scoring using Seurat's AddModuleScore on pucks two and three from sample BRO20 using Lehmann's[21] and Savage's gene signatures, as was done for the Savage *et al.*[1] datasets, which allowed me to identify some of the clusters. Lehmann *et al.*[21]'s TNBC classification is highlighted as

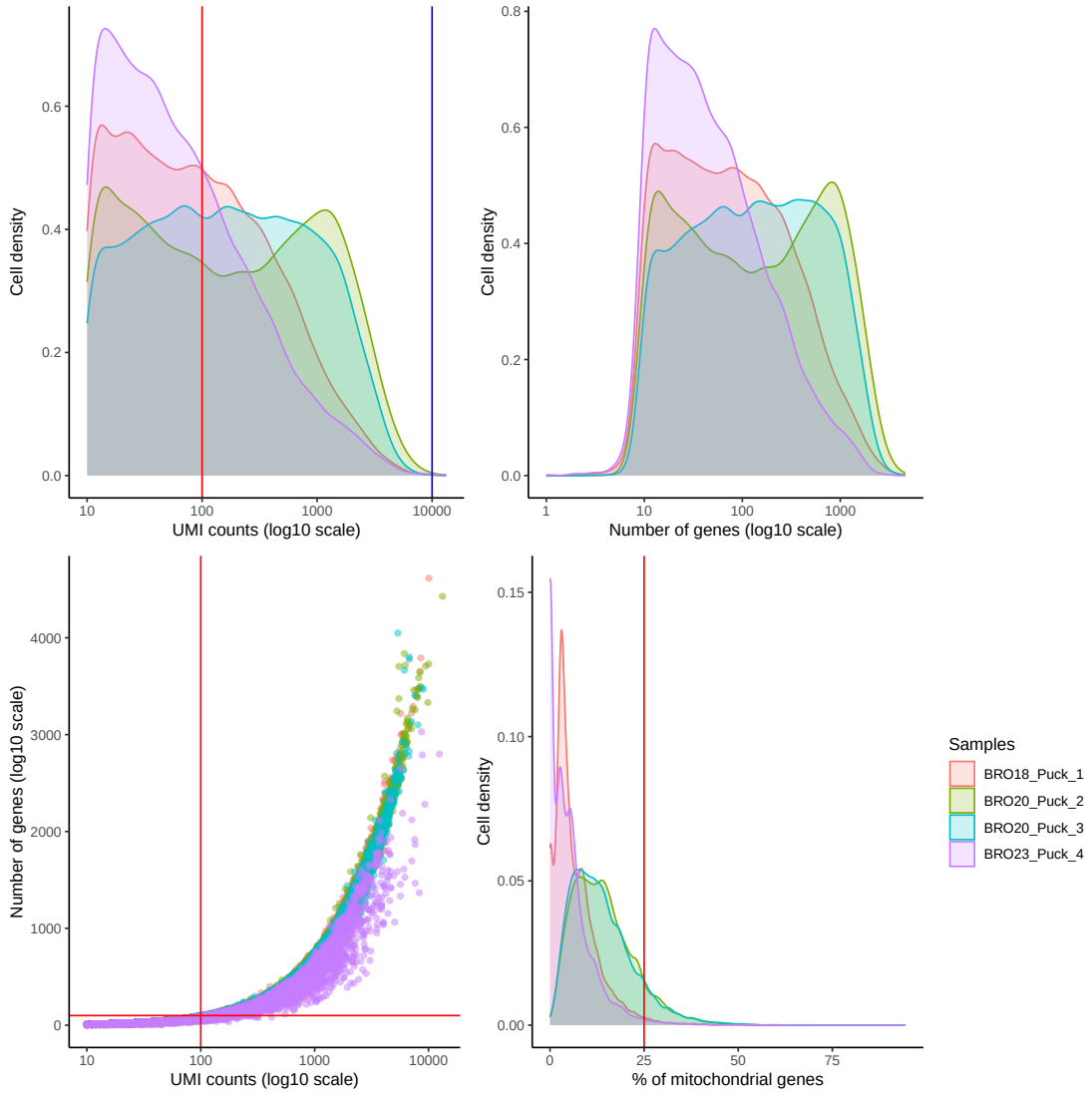


Figure 2.14: Quality control measurements. From top left to bottom right: Distribution of UMI counts, distribution of number of genes, relationship between UMI counts and number of genes, and distribution of the percentage of mitochondrial genes are displayed for every dataset. Red and blue lines indicate filtering thresholds. I generated this figure based on my colleague Clara Pavillet's quality control.

a heatmap for each cluster in figure 2.16 a , while the scoring based on Savage *et al.*[1]'s signature is shown on figure 2.16 b .

Identification of the clusters was challenging probably due to the high number of dropouts in the data, caused by the sparsity and low depth of RNA-sequencing in the technology, which was on its very early stages at the time the spatial sequencing was performed.figure 2.16 b shows how clusters did not score highly for any of Savage *et al.*[1]'s signatures, except for the antigen presentation (which was widespread in puck two). This led me to use Lehmann *et al.*[21]'s classification for scoring and cluster annotation instead, which showed enrichment for the Basal-like type two category in cluster two of puck three, and this puck also showed a distinct signature for the Mesenchymal stem-like category (cluster one). In the case of the Luminal AR signature, virtually the whole of puck three showed some enrichment for it, but it was stronger for cluster zero.

Based on these enrichment scores I annotated puck two in the same way: cluster one as Luminal AR, cluster two as Mesenchymal stem-like, and cluster four as Immunomodulatory. figure 2.16 c shows the spatial location of each cell in BRO20 puck three, colored by the enrichment in Lehmann *et al.*[21]'s gene signatures calculated as detailed above. The UMAP representations of the data can be found in figure 2.17, annotated as described before with Lehmann *et al.*[21]'s TNBC classification.

Finally, only puck three from sample BRO20 was carried forward for *microCJ* analysis due to it presenting a comparatively high number of UMI counts and genes, and the cleanest signatures for cluster annotation, as described above. I believe it provides a good proof of concept of the ability of *microCJ* to import spatial transcriptomics data, which will be demonstrated in chapter 4 and chapter 5.

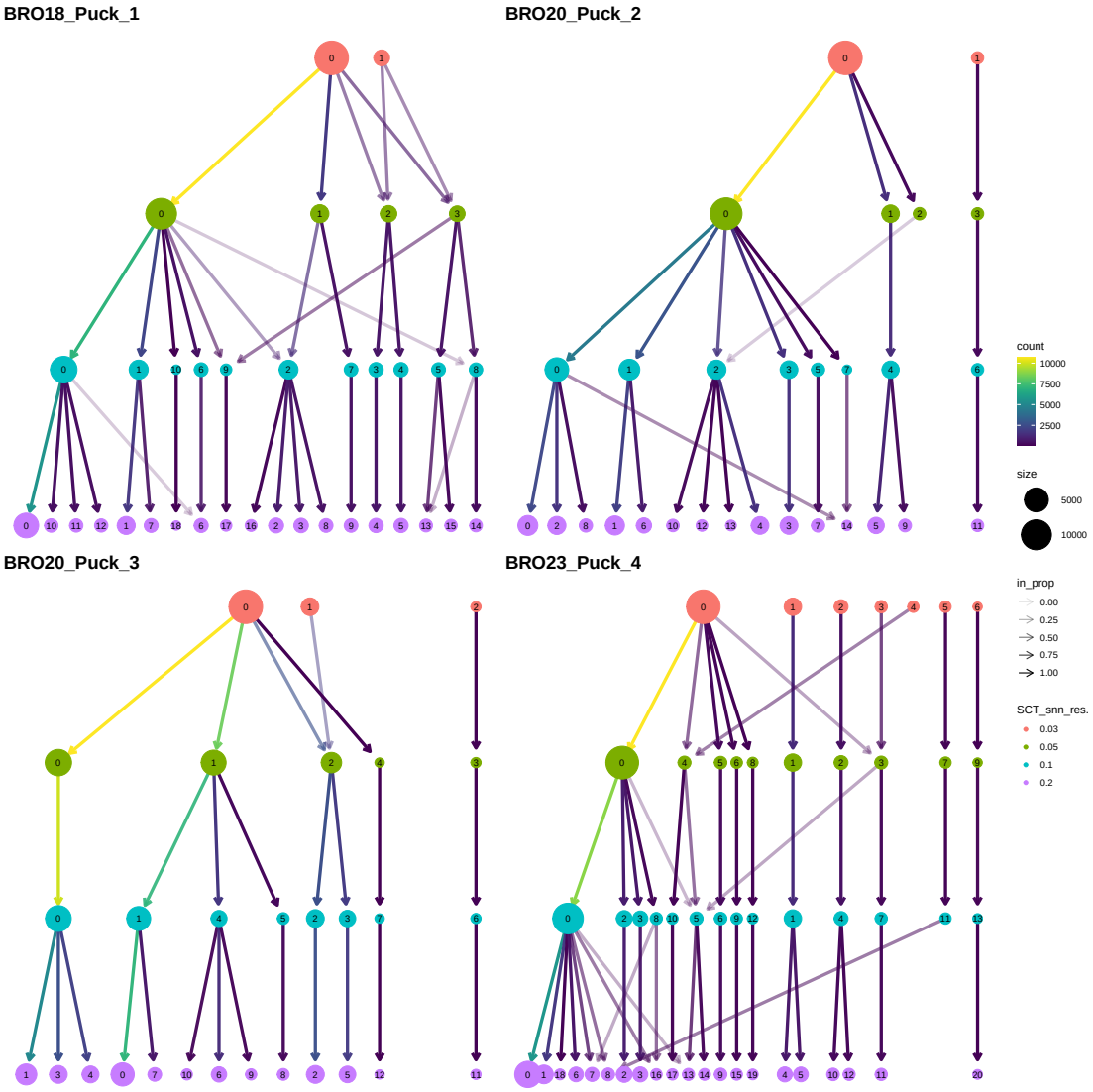
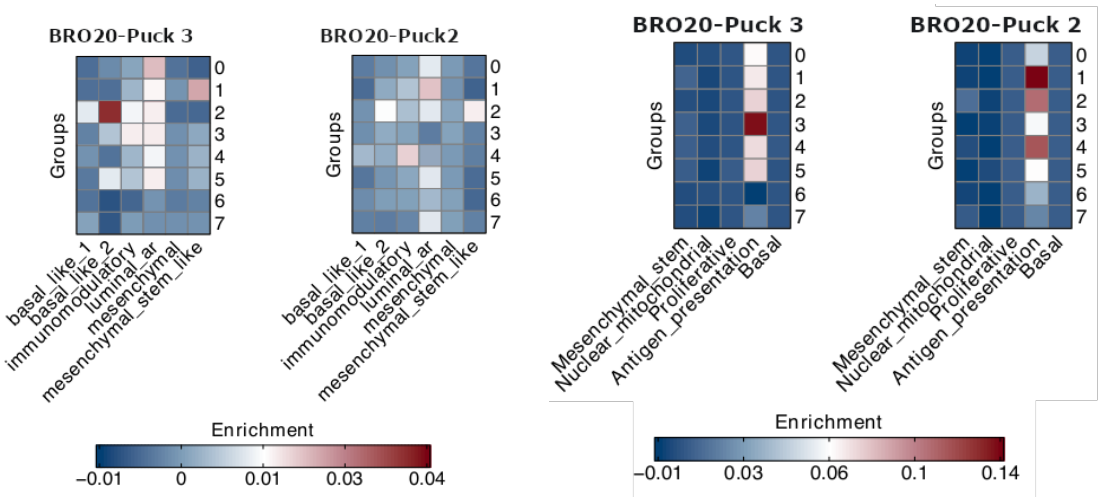
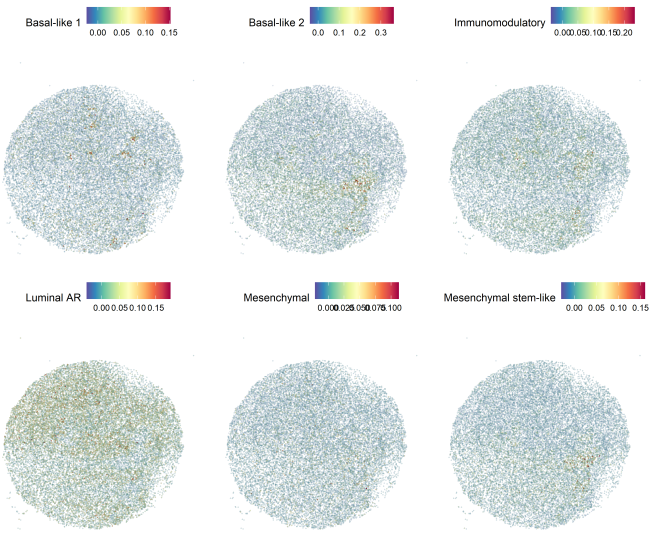


Figure 2.15: Clustering of Slide-seq data. The cluster trees for each sample show how the data is clustered using different resolution parameters.The number of clusters increases with the resolution: stable clusters are maintained with increasing resolution (cells do not split or cross over to far away clusters in the tree). How cells move from one cluster to another as resolution increases is indicated by the arrows. The proportion of cells moving to the new cluster is denoted by the width of the arrow. The size of each cluster is proportional to the circle.



(a) (b)



(c)

Figure 2.16: TNBC subtype scoring. Heatmap and spatial visualization of enrichment in TNBC subtype classifications, with scores calculated by Seurat's AddModuleScore() and then aggregated by cluster, for sample BRO20 pucks 3 and 2, which passed count quality control. (a) shows the enrichment from Lehmann's classification, while (b) shows the enrichment using Savage's classification. (b) shows the actual spatial location of each cell in the sample, colored by AddModuleScore() using genes from Lehmann's classification. Only puck three is represented here, as it showed the highest enrichment compared to puck two, which you can see in the heatmap from (a).

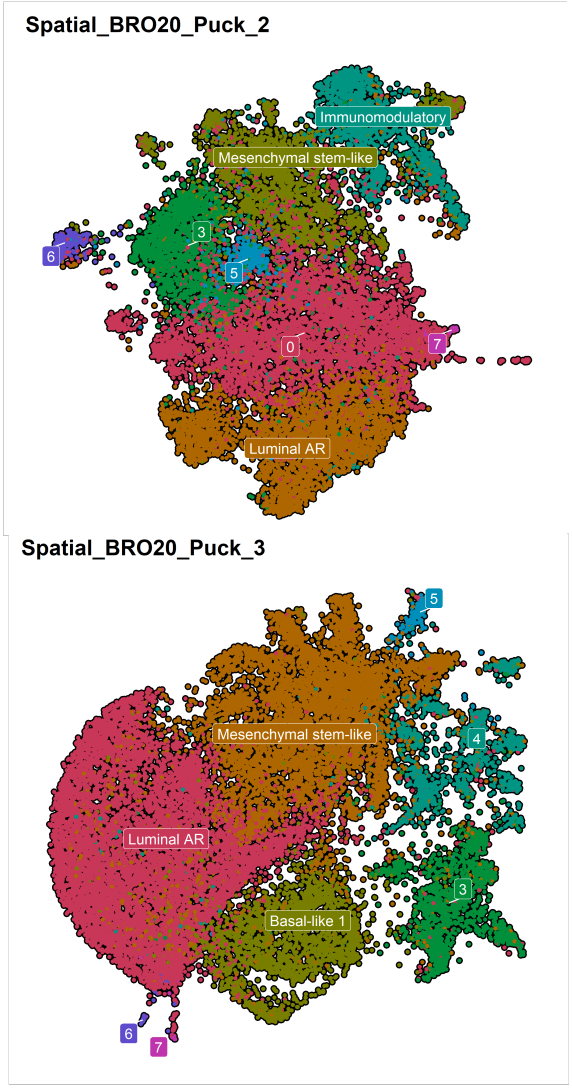


Figure 2.17: UMAP of annotated clusters. UMAP representation of the two Slide-seq pucks (sample BRO20 pucks two and three) that passed the count quality control, coloured by clusters. The biological identity of some of the clusters was identified, and they were consequently labelled as such, leaving the numerical label for the ones with an unclear identity.

2.3 Breast cancer single-cell RNA-seq atlas

Lastly, I used single-cell RNA-seq data from Gambardella *et al.*[297]. In their study, the authors examined 32 different breast cancer cell lines, which together formed a single-cell atlas. They looked at biomarkers expression and found that cell lines are highly heterogeneous in their expression of biomarkers For example, for the seven Luminal B and HER2+ cell lines present in the BC atlas, which are supposed to overexpress HER2 by definition, whereas more than 90% of cells in AU565, BT574 and HCC1954 cell lines express ERBB2, in the other four cell lines ERBB2 expression ranged from 31% of EVSAT cells to 46% of JIMT1 cells and up to 64% of MDA-MB361 cells[297]. They also developed an algorithm to predict response to treatment by linking tumour biopsies to a combination of cell lines, and using the data on those cell lines sensitivity to specific treatments. According to the article, cell lines exhibit transcriptional heterogeneity that allows cells with varying degrees of drug sensitivity to co-exist.

For my analysis, I selected only the TNBC cell lines (type A and B). I then performed quality control filtering: figure 2.18 shows the UMI count, gene abundance and mitochondrial percentage distribution for all selected TNBC samples, and the filtering thresholds selected for the counts (see Methods). Counts were then normalised and scaled, and cells clustered and UMAP dimensionality reduction was performed. The UMAP representation from figure 2.19 shows how each cell line is transcriptionally distinct, each in their own "island". Due to this, I decided to treat each cell line as their own separate cluster and not perform any further sub-clustering.

Figure 2.20 shows the expression level of EGFR in each TNBC cell line. The highest expression, by far, is presented by BT20 and MDA-MB-468, as expected from published literature[125,298]. On the other hand, figure 2.21 does not show, or just very weakly, an enrichment for the EGFR downstream pathway in either of them, which as it happened with the treatment MSL EGFRhi cluster in the Savage *et al.*[1] data, does not necessarily mean that EGFR itself is underexpressed, it could mean that that pathway is not active at that specific point in time, or just

that the EGFR signature is actually not good enough due to incomplete pathway information for example, or technical dropouts in the genes in that pathway.

Figures 2.22 and 2.23 present the enrichment for Lehmann *et al.*[21] TNBC subtype signature and the Savage *et al.*[1] cell type signature, respectively.

I then used the cell-level gene expression matrix from each cell line as another input for *microCJ*. This was an interesting exercise since cell lines are well characterised, and I could compare microCJ results with our biological knowledge (chapter 5).

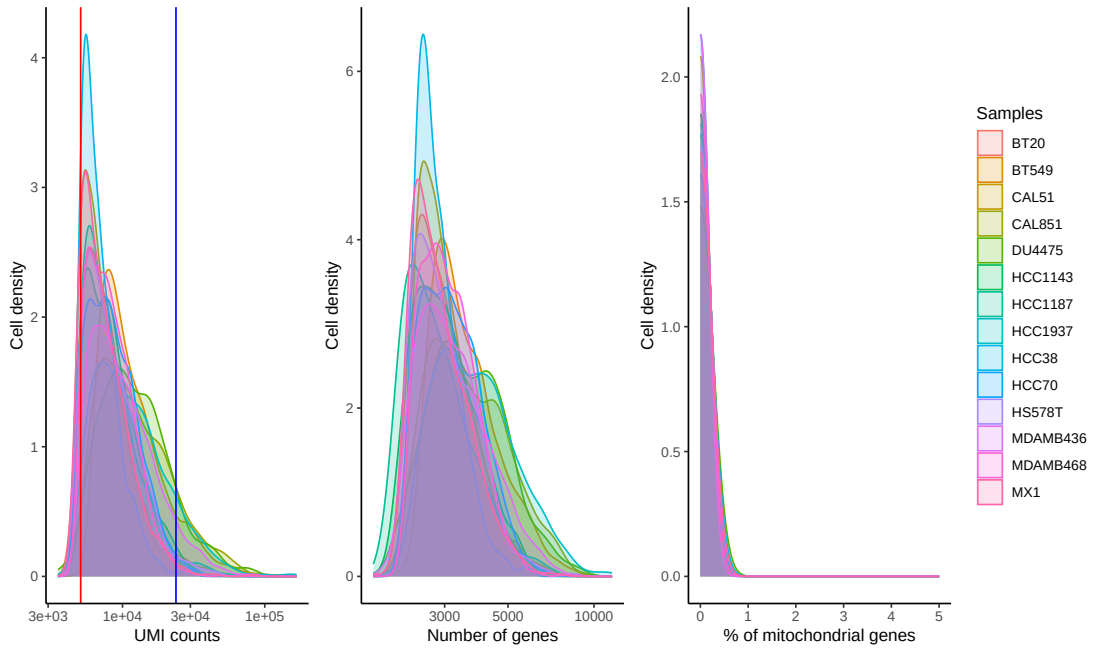


Figure 2.18: Quality control measurements. From left to right: Distribution of UMI counts, distribution of number of genes, and distribution of the percentage of mitochondrial genes are displayed for every cell line. Red and blue lines indicate filtering thresholds.

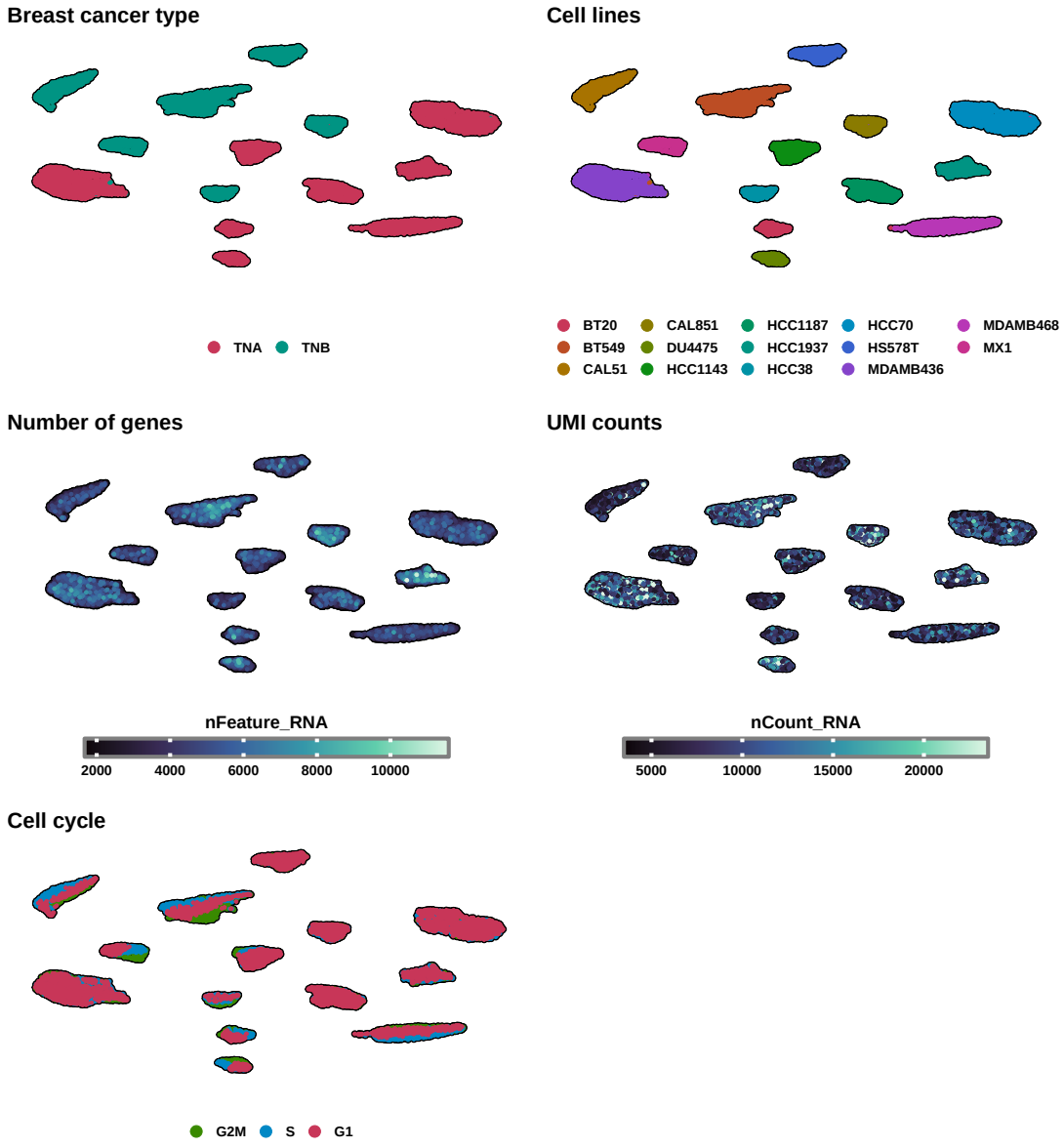


Figure 2.19: Overview of the data used from Gambardella *et al.*[297]. UMAP representation of the data, colored by different features. From top to bottom and left to right: TNBC breast cancer subtypes A and B, cell lines retained, number of genes, gene UMI counts and cell cycle stage.

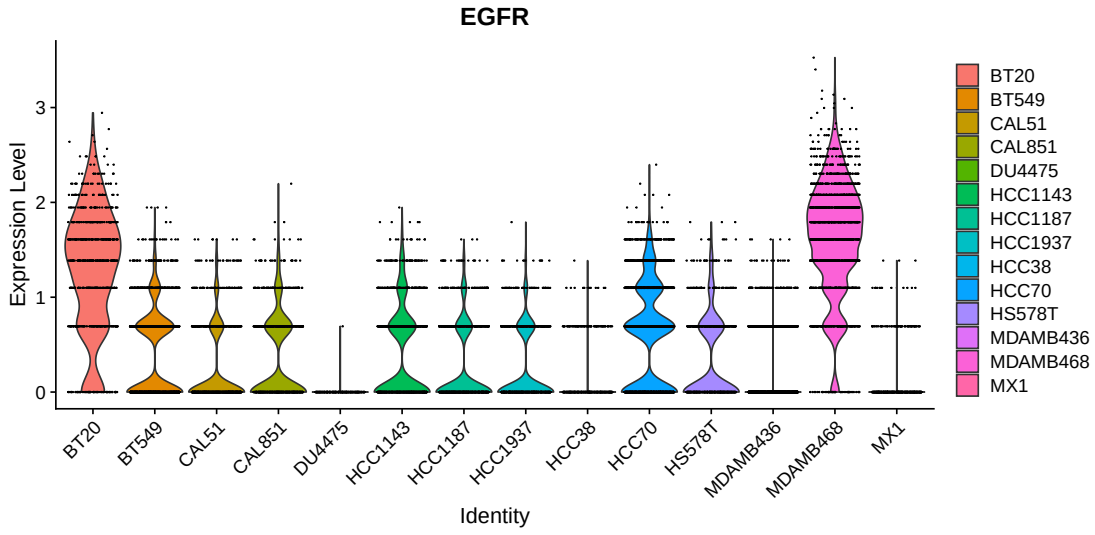


Figure 2.20: EGFR expression. Violin plots of EGFR expression (y axis) in each of the TNBC cell lines (x axis) in Gambardella *et al.*[297] dataset.

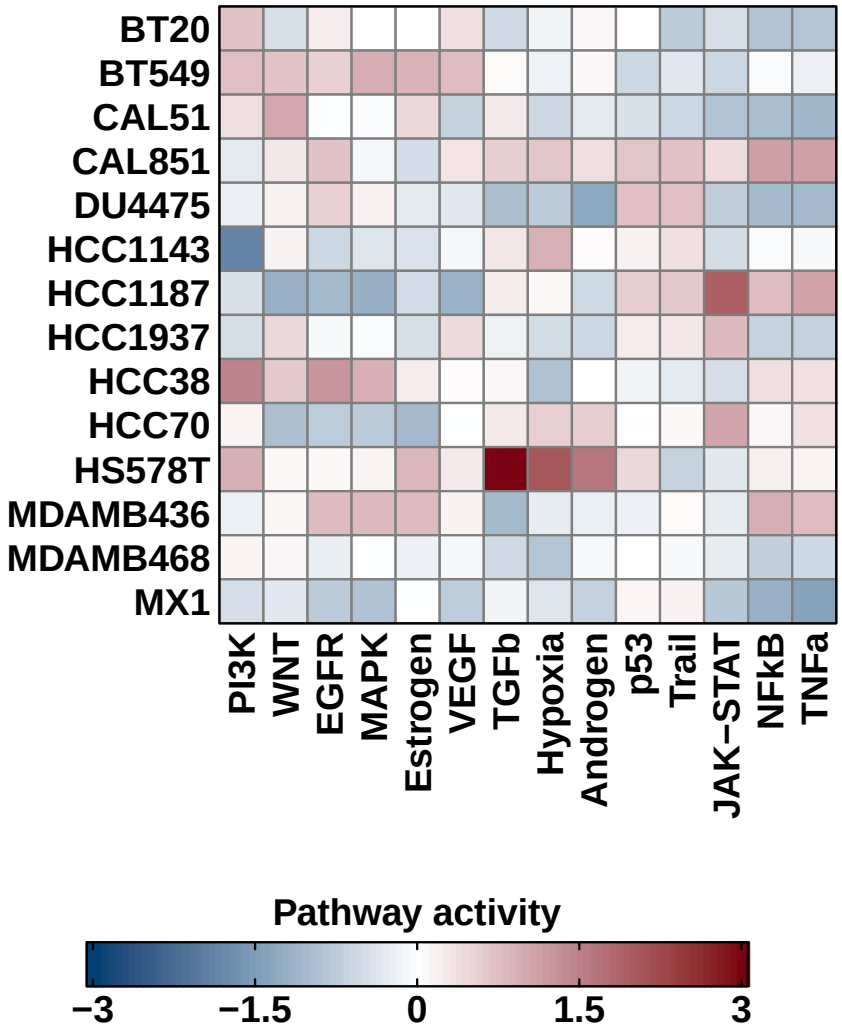


Figure 2.21: Pathway enrichment. Heatmap of the enrichment of several cancer pathways per cell line as inferred by Progeny[294].

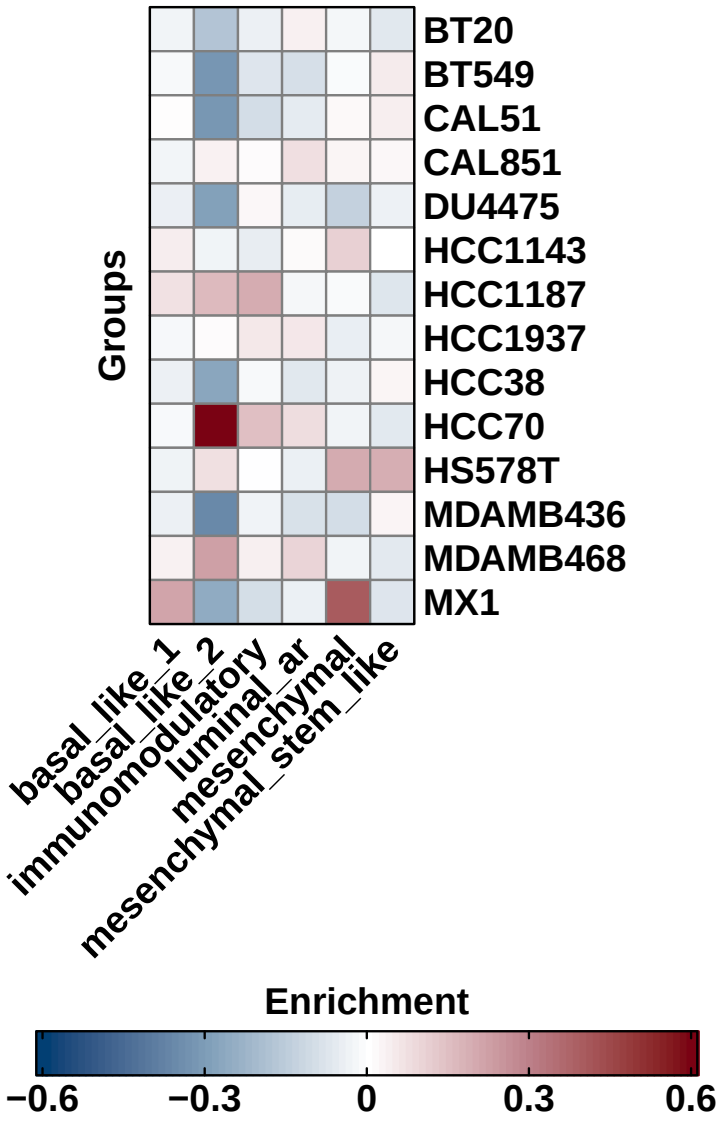


Figure 2.22: TNBC subtype signature for Gambardella *et al.*[297] dataset.. Heatmap showing the enrichment, with scores calculated by Seurat's AddModuleScore() and then aggregated by TNBC cell line, of each TNBC subtype signature (from Lehmann *et al.*[21].

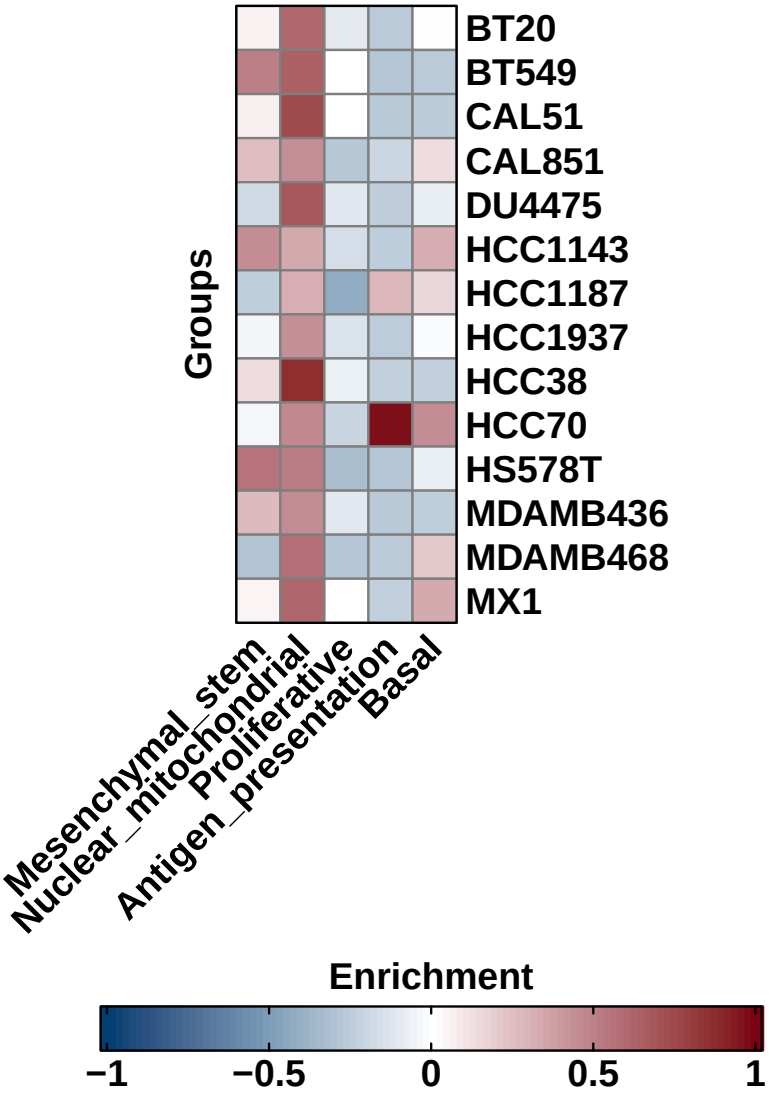


Figure 2.23: Savage *et al.*[1] cell type signature for the Gambardella *et al.*[297] dataset.. Heatmap showing the enrichment,with scores calculated by Seurat's AddModuleScore() and then aggregated by TNBC cell line, of each Savage *et al.*[1] cell type signature.

2.4 Conclusions

In this chapter I have shown the analyses I carried out of different RNA-seq datasets: three single-cell datasets from an exceptional responder to EGFR-targeted therapy (control, treatment and off-treatment), performed by collaborators[1], a collection of four Slide-seq spatial RNA-seq datasets collected by colleagues, and publicly available[297] single-cell datasets from TNBC cell lines. They all provide a experimental basis for my computational model and will serve to demonstrate the ability of *Ingres* (chapter 3) and *microCJ* (chapter 4) to import this data as probabilistic Boolean networks.

I have focused on the exceptional responder datasets[1], as the research they were part of was an attractive exploration of mechanisms of resistance to EGFR-targeted therapy in TNBC. I have managed to reproduce the results of the authors of that research on the control dataset, and have performed similar analyses for the other two datasets, which were not a part of the original article. In general, the results of those analyses agree with the authors hypotheses: a large portion of the surviving cells during treatment do present high levels of EGFR, but are not proliferating. After treatment, most of the cells start progressing through the cell cycle and proliferating again. Still, just from the single-cell data I cannot know for certain that the off-treatment proliferating cells come from the EGFRhi population. Nevertheless, Savage *et al.* [1] demonstrated this using experimental techniques.

Despite this initial corroboration offered by the scRNA-seq data, the mechanisms through which the surviving cells are resisting treatment is still uncleaPror, and more research is needed into this kind of exceptional responder to targeted therapy. The low amount of cells and general noisiness in the RNA-seq data also raises the question of how much can be surmised about this resistance from just the single-cell datasets. If patient samples are still preserved, it would be interesting to repeat the RNA-seq with the higher amount of cells and of reads per cell that come with the current state of sequencing technology; as well as whole genome sequencing to explore the whole landscape of mutations that may be conferring a

growth advantage or a way of resisting EGFR inhibition in resistant cells, however that was beyond the scope of this thesis.

2.5 Methods

2.5.1 Quantification and analysis of single cell RNA-seq from Savage *et al.* [1]

Single cell RNA-seq data from Savage *et al.* [1] was provided by the authors and re-analysed with their permission. Samples were sequenced and mapped to hg19 reference transcriptome as described in their manuscript, including the treatment and off-treatment unpublished samples, using only one run each sequenced with Illumina HiSeq 4000. UMI count matrices generated with CellRanger v. 1.3.1 were imported into R v. 4.0.3 and analysed with Seurat v. 4.2.0[299]. Cells with under 25% mitochondrial genes were retained.

Cell cycle phase was inferred using Seurat's `CellCycleScoring()` using marker genes from Tirosh *et al.* [300]. The transcriptomic differences between cells in the G2M and S phase were regressed, which allowed for the blending of proliferating cells from both phases into common clusters. After normalising and scaling the merged count libraries for the three samples using `SCTransform()`, I clustered cells with the Louvain algorithm using a resolution of 0.6, and performed dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP), following Seurat's standard pipeline. Optimal resolution for cell clustering was visualized using clustree v. 0.5.0[301]. SCTransform returns scaled and centered Pearson residuals ($log(1 + x)$, where x are the sequencing depth-corrected raw counts), and the corrected raw counts themselves. Each of those are used by Seurat at different steps of pipeline and visualization functions.

Marker genes per cluster were found with `FindAllMarkers()`, and gene enrichment analysis was performed using the enrichR package v. 3.1[302] on those that presented positive and significant (Bonferroni-adjusted p-value < 0.01) expression, retaining only terms that presented an overlap of at least five genes. The enrichment for

the Savage *et al.*[1] and Lehmann *et al.*[21] was computed using the Seurat function `AddModuleScore()`. The latter signature was obtained from Karaayvaz *et al.*[295].

Progeny[294] was used to infer enrichment of cancer pathways. For inferring enrichment of transcription factors, decoupleR[303] was employed.

Figures were created with Seurat, ggplot2[304], SCPubr v. 1.0.2[305] and patchwork v. 1.1.2[306].

2.5.2 Quantification and analysis of single cell RNA-seq breast cancer atlas from Gambardella *et al.*[297]

Publically available single cell RNA-seq dat from Gambardella *et al.*[297] was downloaded from `https://figshare.com/articles/dataset/Single_Cell_Breast_Cancer_cell-line_Atlas/15022698`, which is a Seurat merged object with raw counts that had already been prefiltered as described in their publication. Samples were subset to only TNBC cell lines (TNA and TNB), and one with very low number of cells (HDQP1, with fewer than 100 cells) was removed. I performed an additional filtering step, removing cells that presented UMI counts under 5% and over 95% of the overall count distribution, to ensure low quality cells and doublets were appropriately removed. Ensembl gene IDs were converted to gene names using AnnotationHub v. 3.4.0[307] and Ensembl 101 (hg38, `https://www.ensembl.org`). Data were then normalised and scaled, and cells clustered and UMAP calculated as described above for Savage's data. Cell cycle differences were regressed as above.

Progeny[294] and decoupleR[303] were used to infer enrichment of cancer pathways and transcription factors, as above.

2.5.3 Spatial RNA-seq data

2.5.3.1 Collection and preparation of tissue samples

Patients with ER negative, PR negative, HER2 negative invasive breast carcinoma were recruited as part of the BRECO study (full title: "Investigating the relationship between breast cancer cells and the surrounding tissues")[296] and underwent primary surgical resection between January and July 2020 in the Churchill hospital, Oxford,

UK and the Horton General hospital, Banbury, UK. Four clinical tumour samples from three patients (BRO18, BRO20 and BRO23) that had not received previous cancer treatment were included for analysis in this thesis, and were collected with informed consent of the participants in compliance with the National Research Ethics Service South Central Oxford B Research Ethics Committee (project reference19/SC/0025). Four 4 mm tumour punch biopsies (one for patient BRO23, two for patient BRO20 and one for patient BRO18) were collected by Mr. Dennis Remoundos or Miss Toral Gathani for spatial transcriptomic assessment.

2.5.3.2 Library preparation and sequencing

All the following steps in this subsection were performed by Dr Karen Sayal. Breast cancer biopsies were immediately stored in a cryovial with 1 ml of 2-methylbutane (Sigma, catalogue number #M32631) and flash frozen in liquid nitrogen. Samples were stored in a study-compliant manner in a HTA-approved -80 °C freezer between July 2020 and October 2021 in the Department of Oncology, Oxford University. The Slide-seq protocol was completed as per the details shared by the Macosko lab in Boston, USA, as described in their manuscript by Stickels *et al.*[308].

Slide-seq V2 3 mm diameter spatial capture areas ("pucks") were shipped to Oxford from the Macosko lab. 10 µm tissue sections were placed onto the pucks containing 10 µm polystyrene beads linked with a spatially-unique oligonucleotide, synthesised as described by Stickels *et al.*[308]. Tissue permeabilisation, RNA hybridisation, reverse transcription, second strand synthesis, bead clean up and PCR amplification and puck image capture were performed as described in their manuscript. Barcoded libraries were constructed using the Nextera XT kit (Illumina, catalogue number #FC-131-1096) following manufacturer's instructions. A range of custom primers were used as detailed in table 2.1.

Libraries were quantified using the Qubit dsDNA High Sensitivity assay (Invitrogen, catalogue number #Q32854) following manufacturer's instructions. Library quality control was performed on the Agilent High Sensitivity dsDNA chip (catalogue number #5067-4626) using an Agilent 2100 Bioanalyzer by Dr. Sophie Kirschner.

Table 2.1: Slide-seq Custom Primers

Primer Name	Genetic Custom Sequence
Template Switch Oligo	AAGCAGTGGTATCAACGCAGAGTGAATrG+GrG
Truseq PCR Handle	CTACACGACGCTCTTCCGATCT
SMART PCR Primer	AAGCAGTGGTATCAACGCAGAGT
Truseq P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
dN-SMRT	AAGCAGTGGTATCAACGCAGAGTGANNNGGNNNB

After library dilution to 4 nM, they were loaded on the NSQ 500/550 High Output kit v2.5 75 cycles (Illumina, catalogue number #20024906) and sequenced using 75 bp pair-end reads on an Illumina NextSeq 500.

2.5.4 Quantification and analysis of spatial RNA-seq breast cancer samples

Raw sequencing files were processed and mapped by Dr. Evan Murray at the Macosko lab in March 2021, as described in Stickels *et al.*[308] using the slideseq-tools pipeline (`https://github.com/MacoskoLab/slideseq-tools`). Briefly, Picard (`https://broadinstitute.github.io/picard/`) was used first to demultiplex reads, filter out those with low quality, and trim adapters and poliA tails. STAR[309] was used to map reads to the GRCh38.102 reference genome. Finally the pipeline counts the UMIs per gene using the GENCODE hg38 assembly, and matches Illumina barcodes to the puck coordinates. Afterwards, Clara Pavillet processed the four samples with Seurat, performing filtering to cells that retained over 1000 UMI

counts and below 25% percentage of mitochondrial genes, normalising and scaling the data with Seurat's SCTransform(), and performing dimensionality reduction as described above for Savage *et al.*[1]'s data. I re-clustered the cells with the most adequate resolution using clustree, selecting a resolution of 0.1 for the final manual cluster annotation. I then annotated the clusters after investigating the enrichment for relevant gene sets from Lehmann *et al.*[21]'s TNBC classification, using `AddModuleScore()` in Seurat .

How can one capture genes that behave like ghosts,
influencing cells from behind some dark curtain?

— Robert Weinberg

3
Network modelling

Contents

3.0.1 Attractors 90
3.1 The MAPK network 91
3.1.1 Results 94
3.2 The *Ingres* pipeline 105
3.2.1 Implementation 106
3.2.1.1 Infer node activation from RNA sequencing data 106
3.2.1.2 Computing the Probabilistic Boolean Network 107
3.2.1.3 Analysis 110
3.2.2 Demonstration 110
3.3 Conclusions 118
3.4 Methods 119
3.4.1 Network analyses 119
3.4.2 Ingres 120

In this chapter, I will present an examination of the Boolean network (see section 1.2.2 for an explanation of the concepts employed throughout this chapter) that I have used to simulate the intracellular processes of the TNBC cells modelled in *microCJ*. I also present *Ingres*, a computational too to fit Boolean networks with single-cell transcriptomics data. First, I will provide a brief explanation of what attractors are in the context of Boolean networks, which is a key concept for the rest of the chapter.

3.0.1 Attractors

Simulation of Boolean networks can provide insight into the dynamics of biological processes. The identification of recurrent cycles of states, known as attractors, is of particular interest[310]. These are cycles in which the network is trapped as soon as one of its states is reached. Attractors are linked to the phenotypes of the biological context their Boolean network represents, as shown by several studies[310]. Transitions from all states in a Boolean network eventually lead to an attractor, because the number of states in a network is finite. The set of states that lead to a certain attractor is known as its basin of attraction. Biologically relevant attractors often have a large basin of attraction[311,312]. Attractors can be identified simply by iterating state transitions in synchronous and asynchronous networks, following the algorithms described in Hopfensitz *et al.* [310]. There are three types of attractors (figure 3.1):

- **Simple cycles**: loops of states, where any state in the cycle is reached again after a fixed number of successive transitions l, such as $\vec{x}(t) = \vec{x}(t + l)$. This kind of attractor does not appear in asynchronous networks (with one exception described below), because in most cases they have more than one successor state, so that it is possible to leave the cycle.

- **Complex attractors**: overlapping loops that only occur in asynchronous networks, where there possibly is more than one successor state. A complex attractor is a set of states for which all successor and predecessor states are also contained in the set.

- **Steady-state attractors**: attractors that consist of only one state. All transitions from this state result in the state itself. They are the same for synchronous and asynchronous update of a network, and they are a special case of both simple cycles and complex attractors. As simple cycles, they are the exception in asynchronous networks, only happening when the asynchronous update of a network yields only one successor state, which means that in a synchronous update from the first state there is only one gene change.

The behaviour of PBNs is not analysed by identifying attractors, but by using Markov chains to calculate the probability that a certain state is reached after a given number of iterations. States with a non-zero probability of being reached after a high number of iterations are considered analogous to attractors.[216,310].

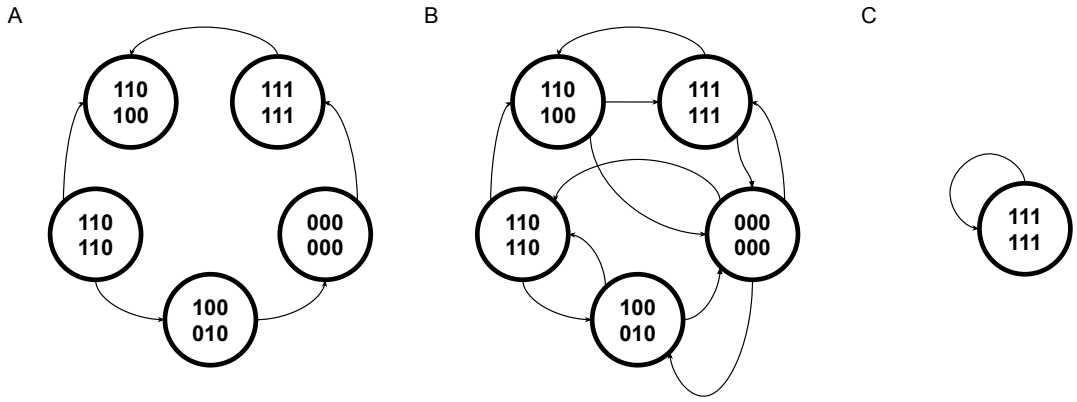


Figure 3.1: Types of attractors in synchronous and asynchronous Boolean networks. A: a simple cycle. B: a complex attractor. C: a steady-state attractor.

3.1 The MAPK network

For the modelling carried out in this thesis I have mainly employed a Boolean network that represents the Mitogen-Activated Protein Kinase (MAPK) regulatory network, as developed by Grieco *et al.*[313] (figure 3.2). The description of *microCJ*, the computational model that is the focus of this thesis, can be found in chapter 4, and its application to TNBC research, in conjunction with the MAPK network, in chapter 5, but in this section I will analyse the network itself in isolation.

MAPK cascades can be activated by a wide variety of stimuli, such as growth factors and environmental stresses. They are implicated in a large array of cellular activities, including gene expression, cell cycle machinery, apoptosis and differentiation[313]. Deregulation of the MAPK cascades are often present in cancer[314,315]. Several components of the network are targets for current or proposed therapy, such as p38, JNK, ERK, MEK, RAF, RAS and DUSP1[313]. The network is focused on the role of MAPK signalling deregulations in tissue homeostasis, the balance between

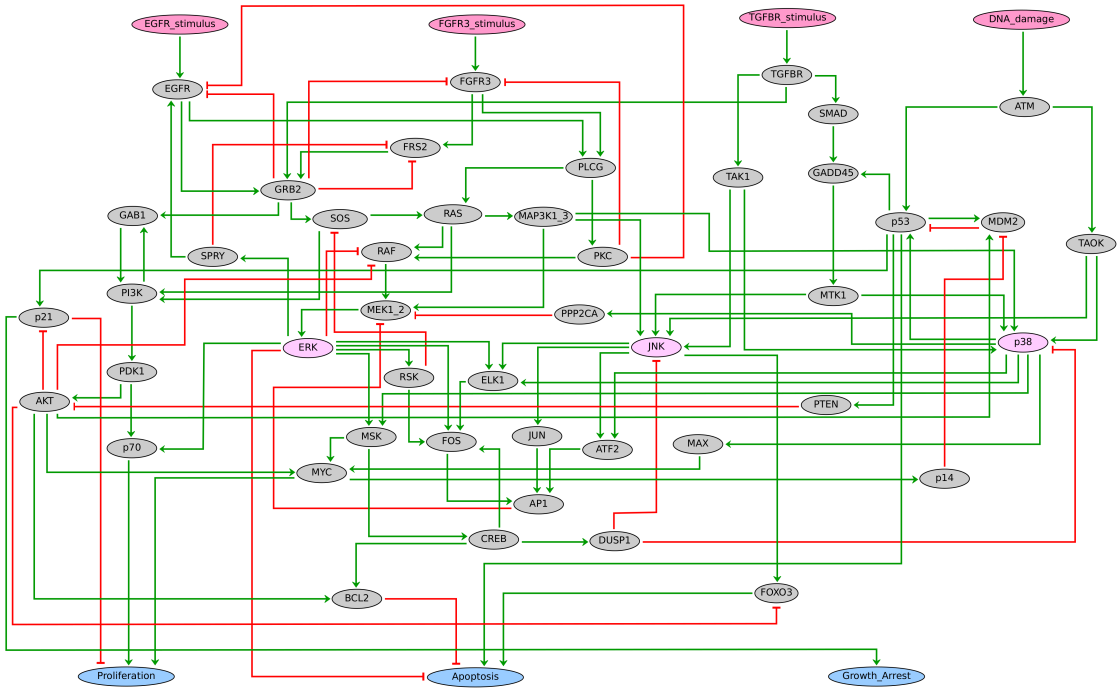


Figure 3.2: The MAPK Boolean network by Grieco *et al.*[313]. Model inputs, phenotypes (cell fates) and MAPK proteins (ERK, p38, JNK) are denoted in magenta, blue and pink, respectively. The full documentation for the network can be found as a supplementary table in the original publication[313]

cell proliferation and cell growth arrest, possibly followed by apoptotic cell death. The choice between these phenotypes, the cell fate decision, is crucial in cancer progression: tumoural cells will receive external and/or autocrine growth stimuli, leading them to cell proliferation and tumour growth, but they will also get anti-proliferative signals, leading to the activation of apoptosis pathways[226]. The goal of the network is to build a predictive dynamical model able to recapitulate the response of the entire MAPK network to selected stimuli, along with its specific contribution to cell fate decision[313]. The whole GRN contains a total of 52 nodes, and it was assembled in a knowledge-driven, mechanistic manner by considering gene and protein interactions as reported in the literature.

The Boolean network then contains three cell fate output nodes: proliferation, apoptosis and growth arrest. The activation of these nodes represent the state of the corresponding processes in the MAPK network, but not necessarily that all requirements for the phenotypes are met, only the ones in which the MAPK

cascades are involved. Specifically, the MAPK GRN considers MYC and p70 in the absence of p21 as markers of cell cycle progression, p21 as a marker of growth arrest, FOXO3 and p53 as markers of apoptosis, and conversely ERK and/or BCL2 indicate apoptosis deactivation.

The MAPK GRN was then extended by Voukantsis *et al.*[257], to be used in the model they built to demonstrate the ABM framework *microC* (figure 3.3). They added a hypoxia responsive module and production of EGF and VEGF. Five new nodes were added. The hypoxia section includes the activation of HIF1, its effect on EGF, VEGF, its interaction with ERK and the necrosis phenotype. This version of the network is the one used by my model *microCJ*, and therefore the one I will consider in this section.

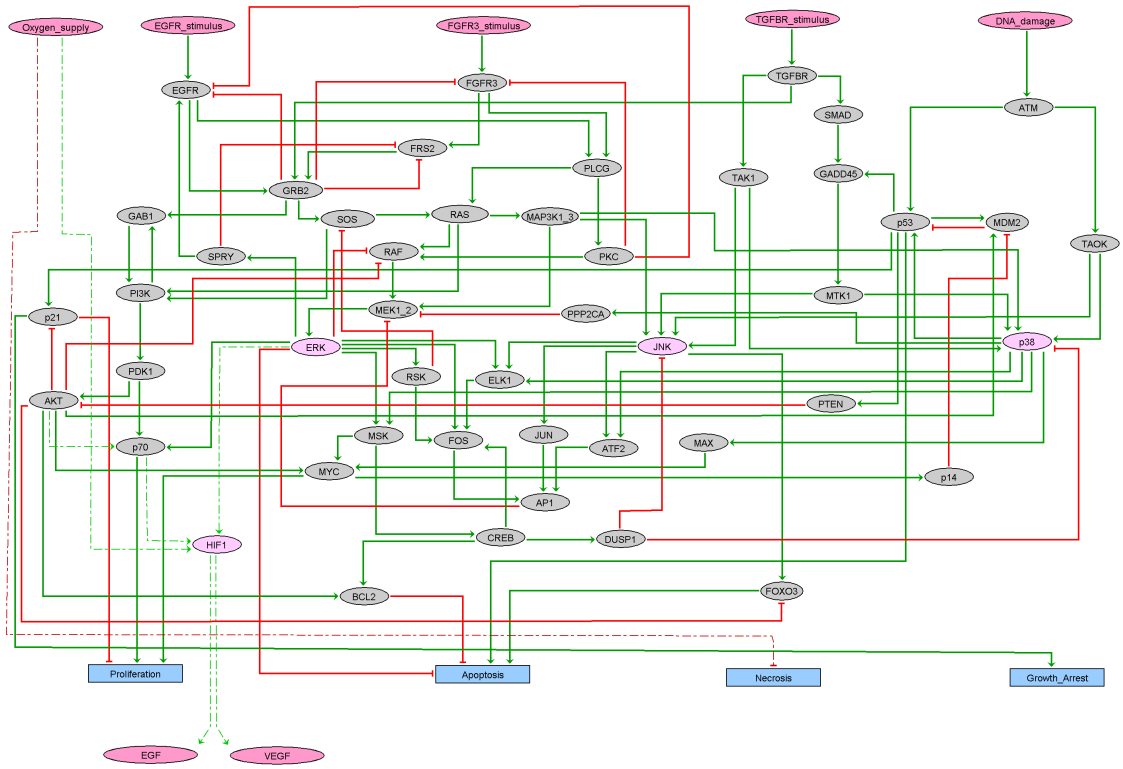


Figure 3.3: The modified MAPK network with added hypoxia module. The dashed lines represent the added interactions. EGF and VEGF are added output nodes, as they do not have outgoing links.

3.1.1 Results

Grieco *et al.*[313] only considered the asynchronous version of their network for the dynamic behaviour analysis presented in their publication, and Voukantsis *et al.*[257] used asynchronous update in their model as well. However, for the modelling presented in this thesis, I have mainly used synchronous networks, because of the shortcomings of asynchronous networks explained above, even though *microCJ* can run any of the three update strategies. Therefore I have tested the dynamics of the network with both synchronous and asynchronous updates, to see if Grieco *et al.*'s[313] analyses of their network hold true for the synchronous version and to perform some additional ones. I perturbed the network mutating — fixing the corresponding node to zero or one for the whole duration of the simulation, which represent mutations that make a gene constitutively active (one) or silenced (zero) — several components of interest, namely p53, EGFR, ERK, JNK, MYC, FOXO3, RAS, AKT, MAP3K1_3, PTEN and PI3K. I used the R package *BoolNet*[316] to carry out these analyses. For some analyses, Grieco *et al.*[313] used reduced versions of the network, due to the computational overhead of dealing with simulating a network that big, with 2^{52} possible states. Yet, I tried different approaches to circumvent this and run my analyses using the full 57-nodes network, because that is what *microCJ* uses.

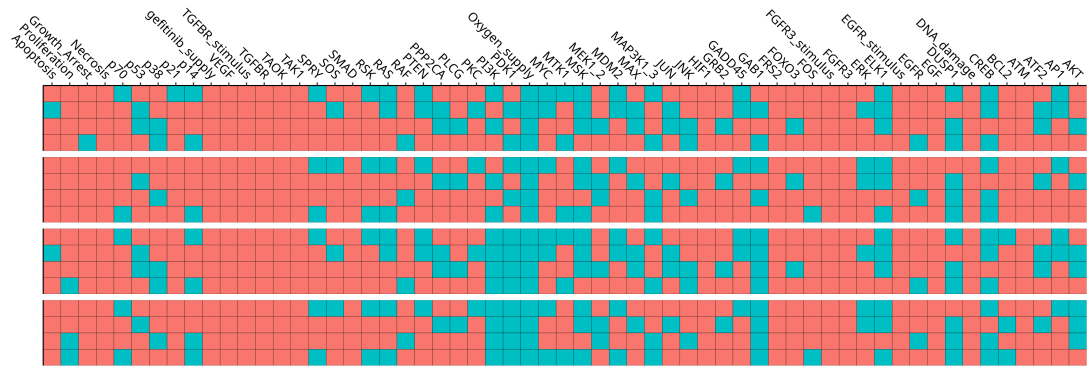


Figure 3.4: Four-state attractors for the unperturbed/inputs-off synchronous MAPK network. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. Each of them is made up of four states that transition to the next one in a loop.

Figure 3.4 shows all the 4-state attractors found for the synchronous network with no perturbations applied, with the exception of all inputs, which are fixed to 0. This would correspond to a "wild-type" model, with no additional mutations. As previously mentioned, we can think of attractors as the "phenotype" of the network under the given set of conditions. It can be seen that with no other perturbations, any of *proliferation*, *growth arrest + apoptosis* or *no activated fate* attractors exist — these nodes are on the left hand side of the plot, and it can be seen that first attractor oscillates between apoptosis and growth arrest, the second attractor does not activate any fate node, the third attractor oscillates between apoptosis and proliferation (which can be interpreted as a lack of a clear decision, or a state where both outcomes are equally likely) and the fourth attractor activates proliferation only. 7-state and 8-state attractors are also found, they are similar to the attractors shown in figure 3.4, with regards to the activated fates. In the case of synchronous networks such as this one, I performed an exhaustive attractor search. Normally attractors are found by starting from a set of initial states, and iterating transitions until a previously visited state is reached, which it is deemed to be part of the basin of attraction of a given attractor. This is repeated until no new attractors appear. If the set of initial states supplied to the algorithm comprises all possible states of the network (2^n states) the search for attractors is exhaustive, that is, all attractors that exist for the network are found. BoolNet uses this methods for networks of up to 29 nodes, above that the number of possible states is too high[310]. Alternatively, an exhaustive synchronous SAT-based search can be used, which formulates the task as a satisfiability problem and solves it using the PicoSAT solver[316,317]. BoolNet uses for this a variant of a method by Dubrova & Teslenko [318]. This is what I have used for attractor search in synchronous networks. The downside for this method is that it does not construct a state transition table. An equivalent exhaustive search could not be performed for the unperturbed asynchronous network, due to the excessive computational power it requires.

Since synchronous networks are deterministic, it follows that the attractor reached from any one state is a pure function of that initial state. To test the state of

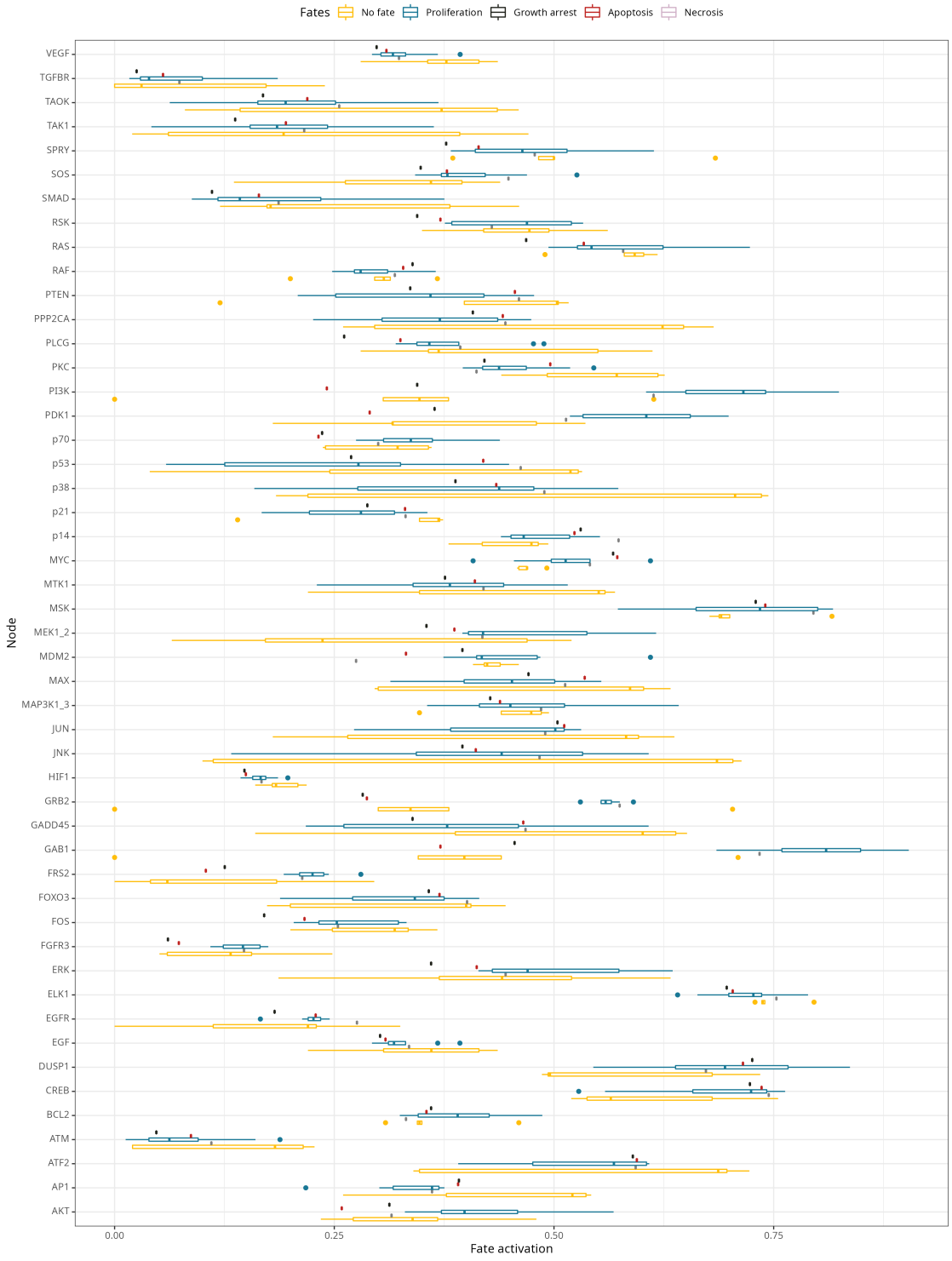


Figure 3.5: Relationship between fate attractors and initial state node activation in the unperturbed-inputs-off synchronous network. For this figure, 10^7 random initial states were created and iterated. For every attractor reached, the average value of the nodes in all initial states conducing to that attractor was computed. The y axis presents every node in the network. The x axis shows the average value of that node for every attractor, separated and coloured by the activated fate in that attractor. If an attractor does not contain an active fate node, it is counted as "no fate". The states are represented as box plots (described in the methods).

which genes has the most influence on which fate attractor is reached, I performed a different analysis. Note that not all attractors are interesting — only the ones that contain an active fate, to which I will refer as 'fate attractors'. These attractors can be biologically interpreted, while an attractor with no active fate nodes represents a cell that has not committed to any fate. As I have mentioned, the SAT-based attractor search does not construct a state transition table, which means that I cannot track what is the initial state that conduced to a given attractor. In order to gather this information, I ran an heuristic search, in which only a subset of the possible states is used, and transitions are carried out until an attractor is reached. I used 10^7 random initial states for this kind of simulation. I classified every initial state by which fate (or lack thereof) is activated in the corresponding attractor. Then, for every node, I noted the frequency of its activation in the initial states conducing to each fate. Figure 3.5 shows the results of this analysis for the unperturbed/inputs-off synchronous network. A total of 16 attractors were found. No gene has a predominant weight on the fate outcome, but their relationship with each fate can be seen. Proliferation attractors tend to be reached when the tumour suppressor p53 is off in the initial state, as it is also the case with TGFBR, which in the MAPK pathway is linked with apoptosis[319]. Again, this kind of simulation is exclusive to synchronous networks, since the asynchronous heuristic search does not construct at state transition table.

After looking at the unperturbed network, I looked at the effect different perturbations to the network had over its attractors, and their relationship with biological reality. First, I simulated EGFR overexpression, this is, I ran the network with the EGFR node fixed to 1, for both the asynchronous and synchronous networks.

Two steady-state attractors were found, identical for the synchronous and asynchronous network, as expected (figure 3.6). Additional 8-state attractors were found for the synchronous network (figure 3.7). The proliferation ones show p53=0, PI3K=1 and AKT or ERK=1; whereas the apoptosis ones show p53=1, and AKT and/or ERK=0. Figure 3.8 shows that the majority of simulations in which the initial state contains p53=0 will reach a proliferation attractor, the opposite is

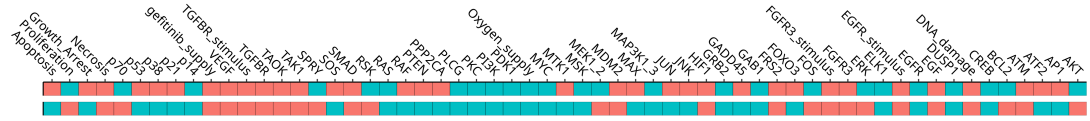


Figure 3.6: The steady-state attractors for over-expressed EGFR. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. There are two attractors, each with one state (known as steady-states).

true for apoptosis: p53 is activated in the initial states that lead to proliferation attractors in fewer than 25% cases, but activated in more than 60% of the cases that lead to apoptosis attractors. The influence of p53 initial state is stronger than that of the other genes, but there are no significant differences.

Then EGFR over-expression was combined with other deregulations, similar to the ones Grieco *et al.*[313] tested. The goal was to see if the modified network managed to reproduce the results reported by Grieco *et al.*[313], and if, furthermore, these results could be reproduced by an asynchronous network. Apart from the steady-state attractors, some other attractors were found, which were almost identical for synchronous and asynchronous update. Figure 3.9 shows the steady-state attractors. It can be observed that most of them make the network reach a single fate attractor. The exceptions are PI3K over-expression (figure 3.9 d), in which the fate reached depend on the state of ERK, AKT and p53 (as it did in the EGFR over-expressed attractors, where PI3K was always active anyway); MEK, which also depends on p53 status; and PTEN loss-of-function(figure 3.9 g), where unless p53 is also off, no fate will be reached. Similar analyses were performed for FGFR3, reaching the same results as Grieco *et al.*[313]. Other biologically relevant facts were tested in the network, some of them are mentioned below.

ERK activation leads to cell proliferation by increasing the transcription of cyclin genes[320]. In the network model, ERK=1 only leads to proliferation attractors (figure 3.10 a). ERK is also reported to disrupt the anti-proliferative effects of TGFβ[314]. In the model, TGFBR=1 leads to apoptotic steady-state (figure 3.10

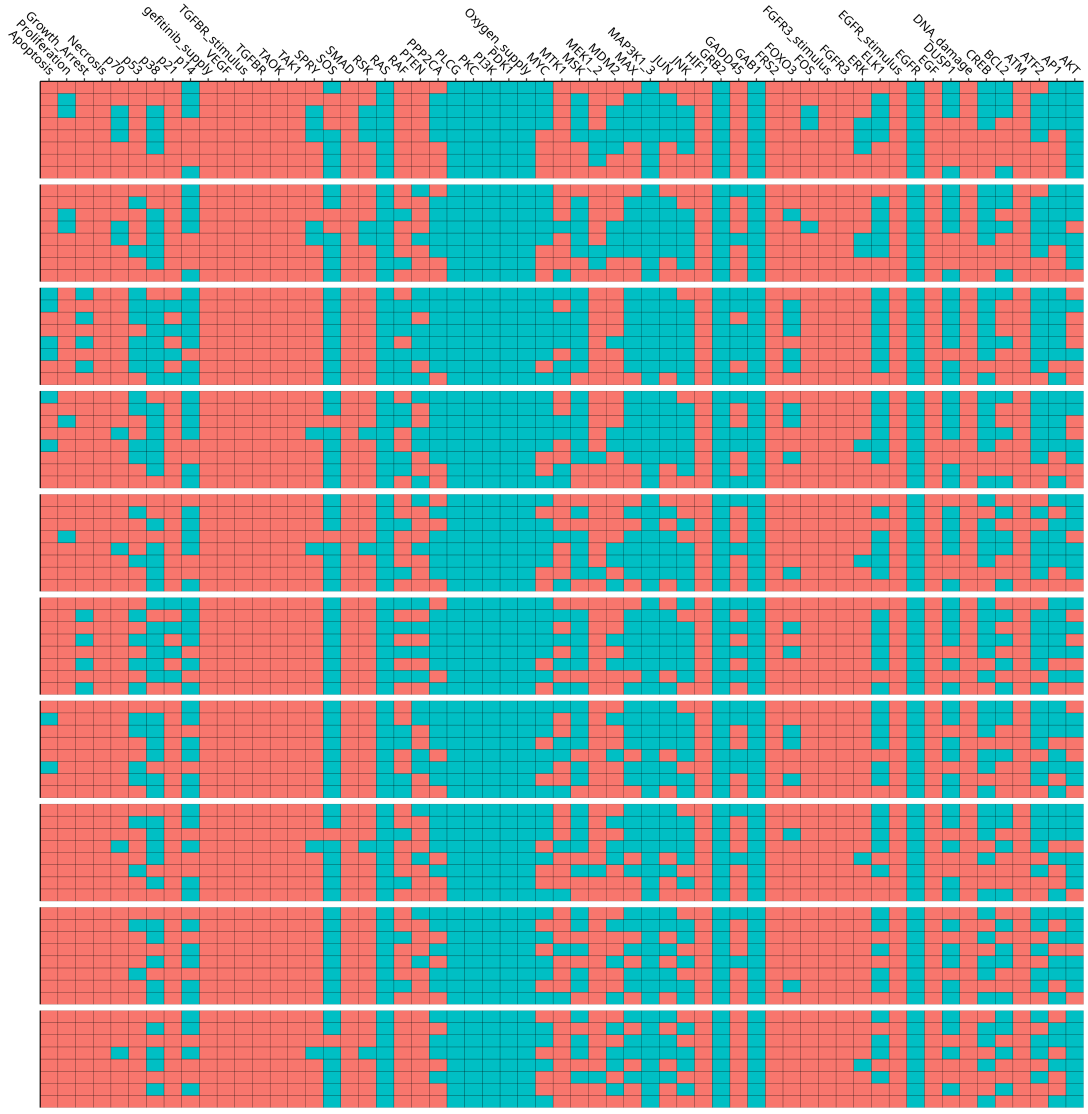


Figure 3.7: Synchronous 8-state attractors for over-expressed EGFR. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. Each of them is made up of eight states that transition to the next one in a loop.

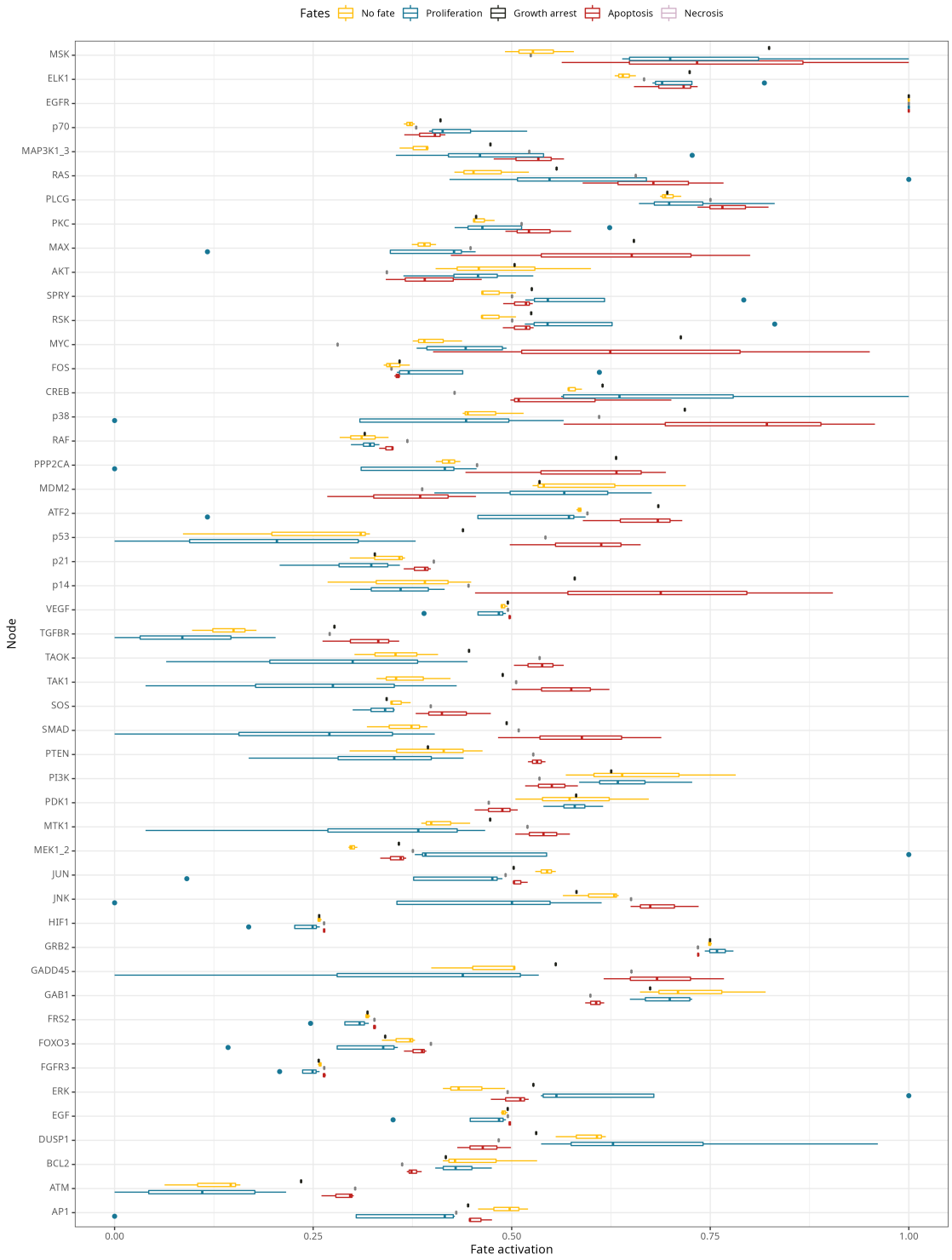
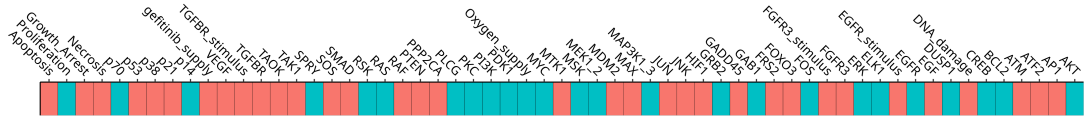
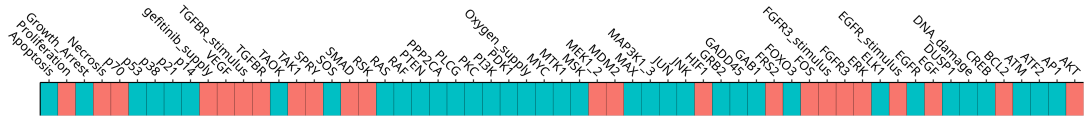


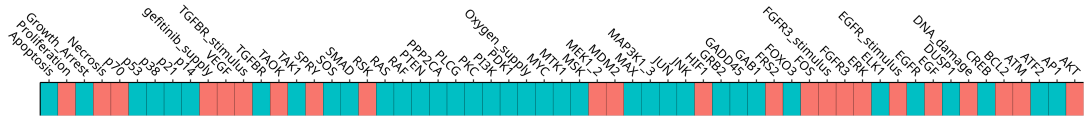
Figure 3.8: Fates and initial states for over-expressed EGFR synchronous network. For explanation of the figure, see figure 3.5 caption



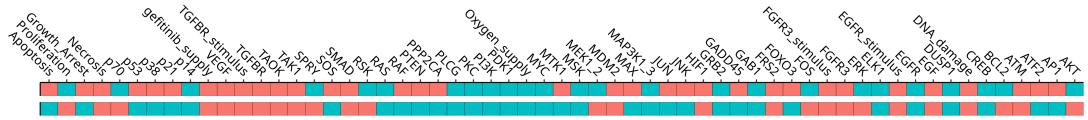
(a) With p53 loss-of-function



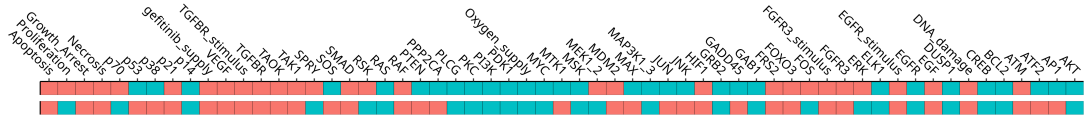
(b) With constant DNA damage



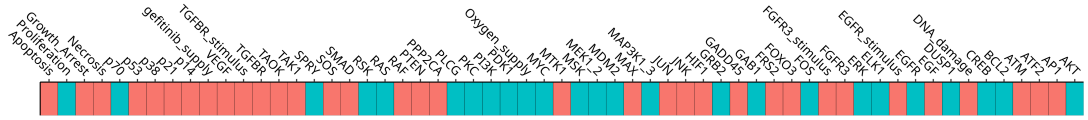
(c) With constant TGFBR input



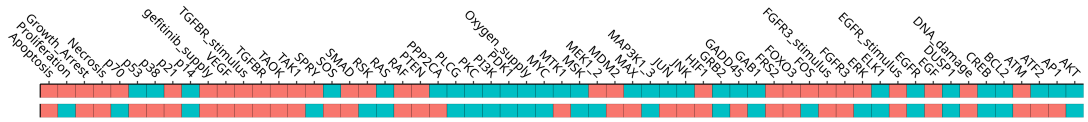
(d) With PI3K over-expression



(e) With AKT over-expression



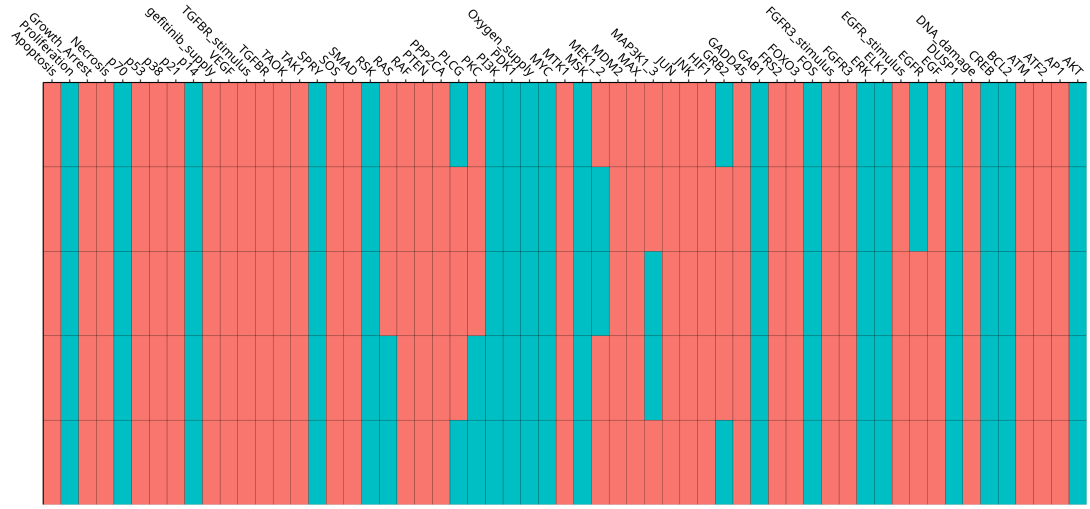
(f) With p14 loss-of-function



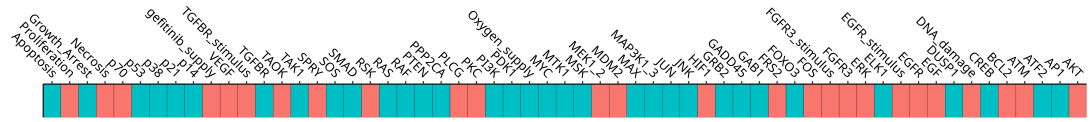
(g) With PTEN loss-of-function

Figure 3.9: The steady-state attractors for over-expressed EGFR networks with additional perturbations. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. There are one or two attractors, each with one state (known as steady-states). a-g represent different perturbations.

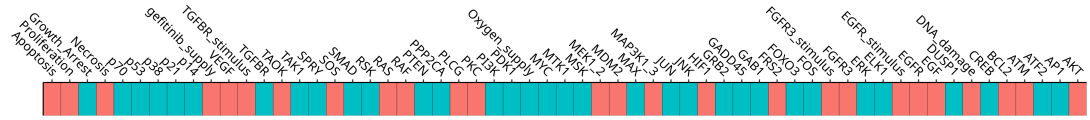
b), but coupled with ERK=1 leads to a growth arrest steady-state instead (figure 3.10 c).



(a) ERK over-expression



(b) TGFBR constant input



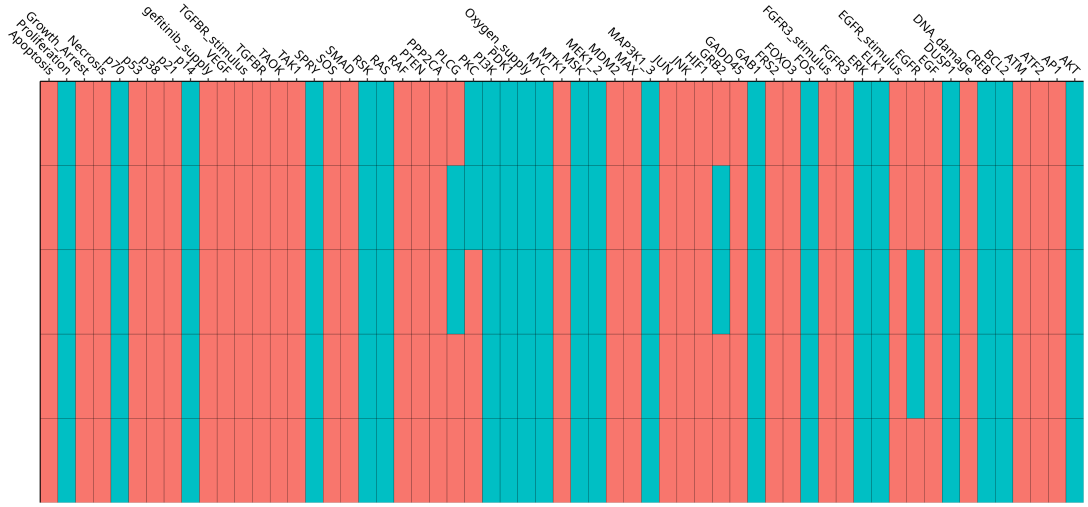
(c) ERK over-expression in the presence of TGFBR

Figure 3.10: ERK disrupts the anti-proliferative effects of TGFβ. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. a-c represent different perturbations. There is one attractor in each of a-c: a) with five states, and b-c with one state (known as steady-states).

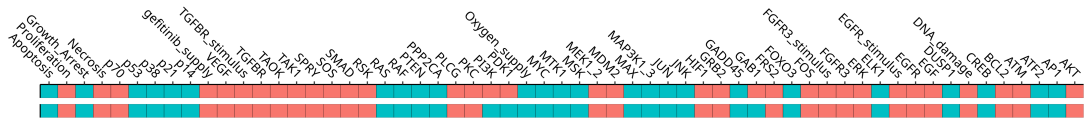
JNK might inhibit RAS-dependent tumour formation and promote apoptosis[321]. In the model RAS over-expression leads to proliferation attractors (figure 3.11 a), but only if JNK is not active, otherwise an apoptosis attractor is always reached (figure 3.11 b).

In conclusion, the MAPK network has proven to be robust and to return mechanistically interpretable results under a variety of perturbations, despite its small

scope: indeed, the reported results by Grieco *et al.*[313] can be reproduced in asynchronous networks, despite the added level of stochasticity. This small scope is a necessity when dealing with mechanistic models, as explained in section 1.2.1 (page 32). The next steps for this network will be to build on top of it, to adapt it to my research focus and provide better explanations for EGFR pathways and resistance strategies, as it will be shown in chapter 5 (page 195).



(a) RAS over-expression leading to proliferation



(b) RAS and JNK over-expression

Figure 3.11: RAS-mediated proliferation only occurs in the absence of JNK. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. a-b represent different perturbations. There is one attractor in a), with five states. There are two attractors in b), with one state each (known as steady-states).

3.2 The *Ingres* pipeline

For the reasons stated above, throughout this thesis I have employed PBNs to better capture the complexity of single-cell transcriptomics data, as produced by my *Ingres* pipeline. *Ingres* is a computational workflow, released as an R package, that, given single-cell transcriptomics data and a Boolean network as inputs, produces a PBN for every single cell in the data, or for every cluster of cells.

A limitation of Boolean network modelling, which I also address with *Ingres*, is the simulation of continuous biological variables. For example, different levels of gene expression between cell sub-populations, or in cells exposed to different conditions, cannot be modelled in Boolean networks. In these, over- and under-expression are usually modelled by imposing a constraint to the node, fixing it to a single state throughout the simulation. The Boolean function for those nodes will be ignored and its state will always be one or zero, regardless of the state of connected nodes. Of course, this is the trade-off of discretisation, which is normally thought to be offset by the advantages it provides: an easier and faster computational analysis, homogenisation of different data sets and reduction of noise[322].

On the other hand, quantitative and hybrid models have been proposed to represent biological networks, and they allow the simulation of continuous levels of expression, but they are only feasible for very small networks, since they require an extensive knowledge and experimental data to be developed[323].

Ingres provides another solution to this problem by representing different levels of activation/expression while still working with Boolean functions. This enables to fit the PBN using single-cell RNAseq data, which would be otherwise not possible. Specifically, it uses a previously developed algorithm, VIPER[324], to infer protein activity starting from a gene expression matrix and a list of regulons —defined as the list of transcriptional targets of each protein. This computes a matrix of normalised enrichment scores (NES) for each protein. The expression of these regulons represent an optimal indirect method to measure the activity of a specific gene/protein (as represented by our nodes in the network).

BGRNs are also increasingly used in combination within more complex, multi-scale models, such as my own *microCJ* (chapter 4). These are models where each cell (e.g. modeled as an independent entity using Agent Based Modelling, see section 1.2.3.4) is simulated as having their own gene network whose output governs the behaviour of the cell. External signals, nutrients or interactions with other cells can also be received by the cell, and act as inputs of its inner network, therefore simulating tumour dynamics at several levels: the microenvironment level, the cellular level with its cell-cell interactions and the intracellular level. These multi-scale models are a novel and exciting development in the field, and there are multiple interesting examples such as the ones produced using the modelling frameworks *microC*[257] or PhysiBoSS[325]. For this kind of models, finding a way to represent the differences between cells and between subpopulations can be of interest, especially for cases where cellular heterogeneity and competition have a significant role in driving the biology of the system[326].

I hope that *Ingres* will help bridge the gap between highly dimensional genomic/-transcriptomic dataset and discrete computational models, providing a useful tool for generating mechanistic explanations and identifying actionable targets for therapy, tasks for which these models are particularly suited.

3.2.1 Implementation

First, I will describe the steps in the *Ingres* workflow, then I will provide some examples and results. I implemented *Ingres* as an R package[327], and the documentation attached to it contains specific information about the functions that are used to run each of the steps described below.

3.2.1.1 Infer node activation from RNA sequencing data

This step is crucial to retain some of the information that is necessarily lost not only with the discretisation of the data, but also by retaining only the expression scores for the few dozen genes contained in a GRN. Since VIPER takes into account all genes that are targets of the proteins of interest —the ones contained in the

GRN— in order to generate the Normalised Enrichment Score (NES), using NES instead of gene normalised expression preserves more information about the protein of interest, and therefore makes the resultant model more representative of the biological reality that is being studied. Using these scores instead of simply using the RNA levels also addresses the key issue that RNA levels are often not a good proxy for protein activity, which is often ignored in network studies.

VIPER uses an algorithm called *aREA* (analytic Rank-based Enrichment Analysis) to infer protein activity by analysing the expression of the protein's regulon (the list of transcriptional targets of each protein), which is tissue-dependant[324].

VIPER probabilistically integrates the target mode of regulation (activation or repression), statistical confidence in regulator-target interactions and target overlap between different regulators[324]. VIPER is also very robust and resilient to signature noise and to incomplete regulons[324]. *Ingres* runs VIPER on the provided gene expression matrix, using the regulon supplied by the user. This regulon can come from the ARACNe algorithm[328,329] or from other databases such as Dorothea[330].

If several regulons are provided (as it occurs in single cell RNA-seq data), the metaVIPER algorithm[331] will be run instead. This algorithm is designed to integrate multiple interactomes, not necessarily tissue-matched (this could be an advantage over VIPER, which does need tissue matched regulons). This is most valuable in single-cell RNA-seq data, due to its heterogeneity, inherent noisiness and low sequencing depth[331]. Indeed, metaVIPER has been shown to reduce bias and batch effect, generating reproducible protein activity signatures[331]. Therefore, this is the recommended method for *Ingres*. Collections of tumour regulons are available at Bioconductor[332].

3.2.1.2 Computing the Probabilistic Boolean Network

Ingres takes $\vec{s}$, the vector containing all the NES generated by VIPER for a given cell or cluster of cells, and an user defined range such as

$$D = \{d_1, d_2\},\ d_1 + d_2 = 0,\ d_2 > d_1,\ d_2 \leq 1$$

to which the scores are to be rescaled, and computes the rescaled scores (r) with mid-point zero as

$$r_i = \frac{s_i(d_2 - d_1)}{max|\vec{s}|}$$

for every element of $\vec{s}$, where $max|\vec{s}|$ is the maximum absolute value through the NES vector. Then, in the new PBN that *Ingres* produces, to every node x_i where $r_i \neq 0$, there corresponds a set $F_i = \{g_i, o_i\}$ where g_i is the node function in the original Boolean network and

$$o_i = f_i(r_i) = \begin{cases} 1 & r > 0 \\ 0 & r < 0 \end{cases}, f_i : \mathbb{R} \setminus \{0\} \to \mathbb{B}$$

For nodes x_i where $r_i = 0$, $F_i = \{g_i\}$. For every set F_i there is a corresponding set of probabilities $C_i = \{c_i^g, c_i^o\}$, where

$$c_i^o = |r_i|,$$
$$c_i^g = 1 - c_i^o$$

with c_i^o and c_i^g being the probabilities for the functions o_i and g_i, respectively. For F_i with only one function g_i (when $r_i = 0$), $c_i^g = 1$. Those probabilities will be used to determine which function should be use to perform every transition $x_i(t) \mapsto x_i(t+1)$.

If the PBN is to be computed for a given cluster (a subpopulation of cells), the median NES among all cells belonging to that cluster is calculated for each gene, and then the steps above are executed taking the set of all cluster scores computed this way.

I believe this is a good middle point between the large, continuous and noisy raw expression data, and the rough discretisation approach of representing over-expression by fixing nodes to one and vice versa. Nodes will still have one of two values, active/inactive, but they will not be fixed into a single state — except for the top expressed gene in each cell, which will have a fixed function probability of 1, they will still be subject to their original Boolean function, but they will

have a bias towards inactivation or activation directly proportional to their protein activity enrichment score (Figure 3.12). This means that a node with probability of 1 will be five times as likely to be active as a node with a probability of 0.2. This bias will of course work independently from the Boolean function: at any update step where the fixed function is chosen, the incoming edges of the node will be ignored and the state will be determined by the fixed function. This simulates the action of all the factors that contribute to driving the phenotype but are not present in the GRN: other genes and proteins, external signals, environmental factors, etc., without a significant increase in complexity other than what naturally comes with switching to a probabilistic network. The user, through the rescaling range parameter (which defaults to [-1, 1], the maximum), can decide how strong this bias can generally be in the generated network.

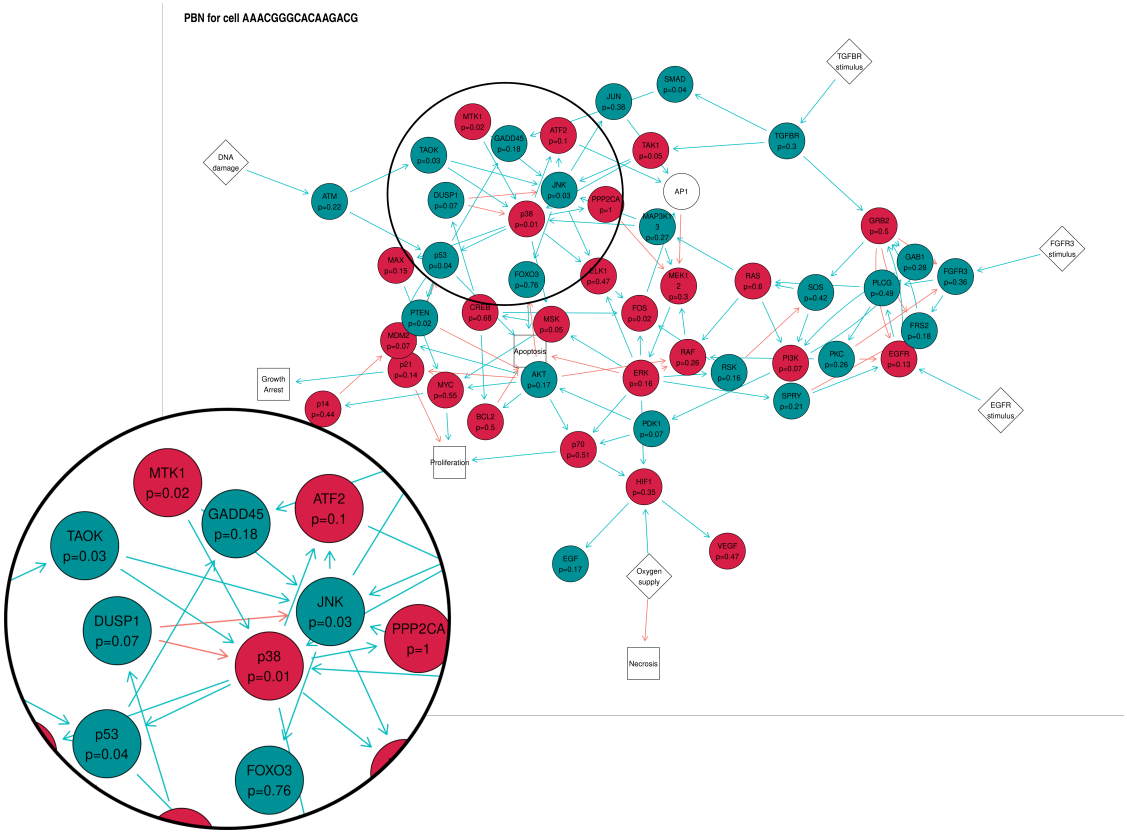


Figure 3.12: A network generated by *Ingres*. A network plot showing the probabilities for the fixed function of each node. The green nodes indicate that their fixed function is one (activation) and for the red nodes it is zero (deactivation).

3.2.1.3 Analysis

Now, the user can explore the results by using *Ingres* functions to plot the network for any cell or cluster. The results can be plotted as heatmaps too. Additionally, an interactive app can be run, where the user can see all cells in a tSNE plot — by providing a Seurat object[333] — and click on any of them to be shown its corresponding PBN (Figure 3.13).

Finally, *Ingres* provides several wrapper functions for relevant parts of *BoolNet*[316], which can be used to perform analyses on any PBN produced by *Ingres*, such as computing its attractors[310]. This can be useful to identify predicted phenotypes of the modelled network. *Ingres* exports its networks in *tidygraph*[334] and *BoolNet*[316] format, so they can be further interrogated using other software, such as *GinSim*[335].

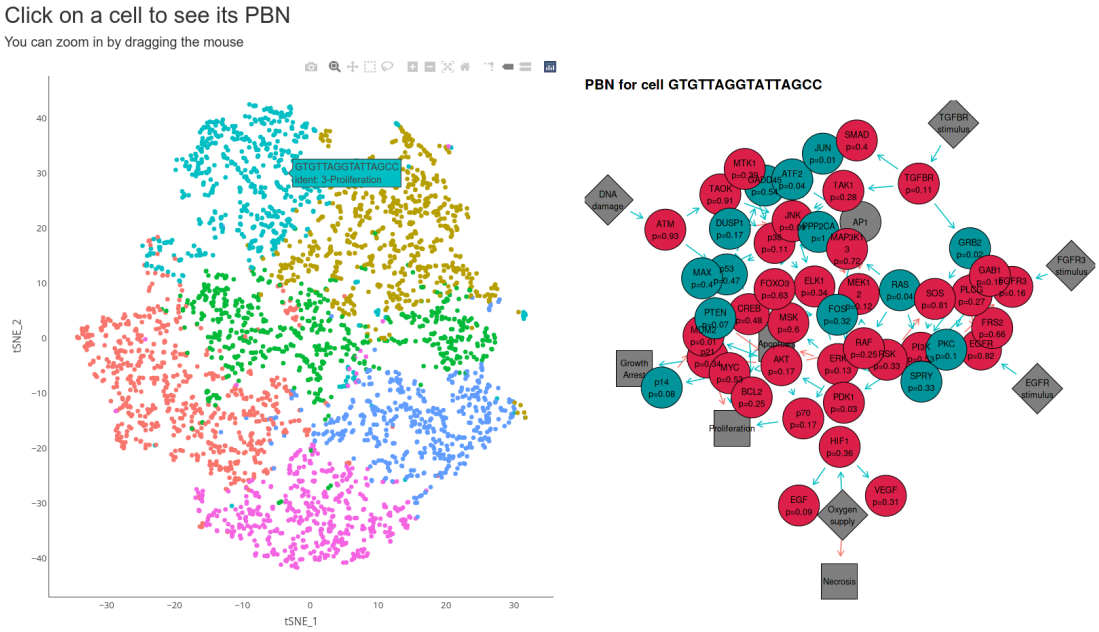


Figure 3.13: Interactive app to explore the results. Screenshot of the app that allows the user to click on any cell on the tSNE cell and shows the computed PBN of the selected cell.

3.2.2 Demonstration

As a means to test *Ingres* independently from *microCJ*, I compared the behaviour of some PBNs produced by it with the expected phenotype. I used a growth

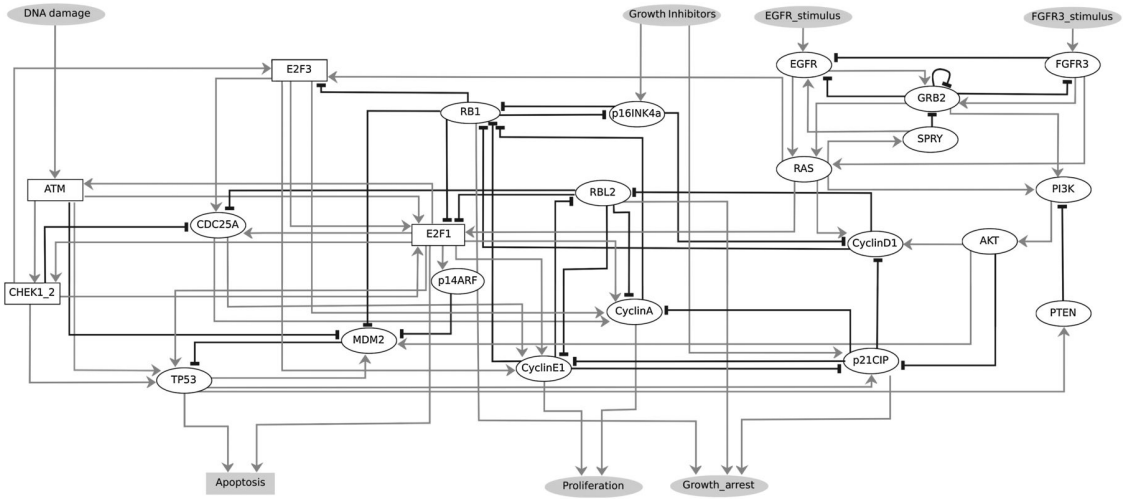


Figure 3.14: The network of the components involved in entry into apoptosis and cell cycle, through E2F1, by Remy *et al.* [336]. Gray arrows, positive influences, T-shaped arrows, negative influences

factor and cell cycle signalling network for bladder cancer found in the literature[336] (figure 3.14), which after some changes to remove redundant nodes and make it smaller consisted of 30 nodes. I chose a recently published single-cell RNA-seq atlas of 32 breast cancer cell lines[297] as the input for *Ingres*. For every cell line, I clustered cells based on their cell-cycle score, as described by the Seurat authors[337], in three clusters, G1, S and G2M. I performed the whole *Ingres* pipeline on each dataset, obtaining a PBN for each of the three clusters. Finally, I computed 1000 state transitions for each network to see what the fate activation trend is for each of them (Figures 3.15 and 3.16). As expected, as the cell cycle progresses, the networks corresponding to its phases show a increasing proportion of proliferation, and in the G2M network, this is higher than apoptosis. This result is very robust across the different breast cancer cell lines in the input data (figure 3.15). Therefore, as long as the input network is chosen based on the phenotype being studied, the PBN generated by *Ingres* preserves the expected behaviour of each of the cell clusters.

The same simulation using a bladder cancer dataset[338] (available as GSE130001) as the input for *Ingres* was performed with similar results (figures 3.17 and 3.18).

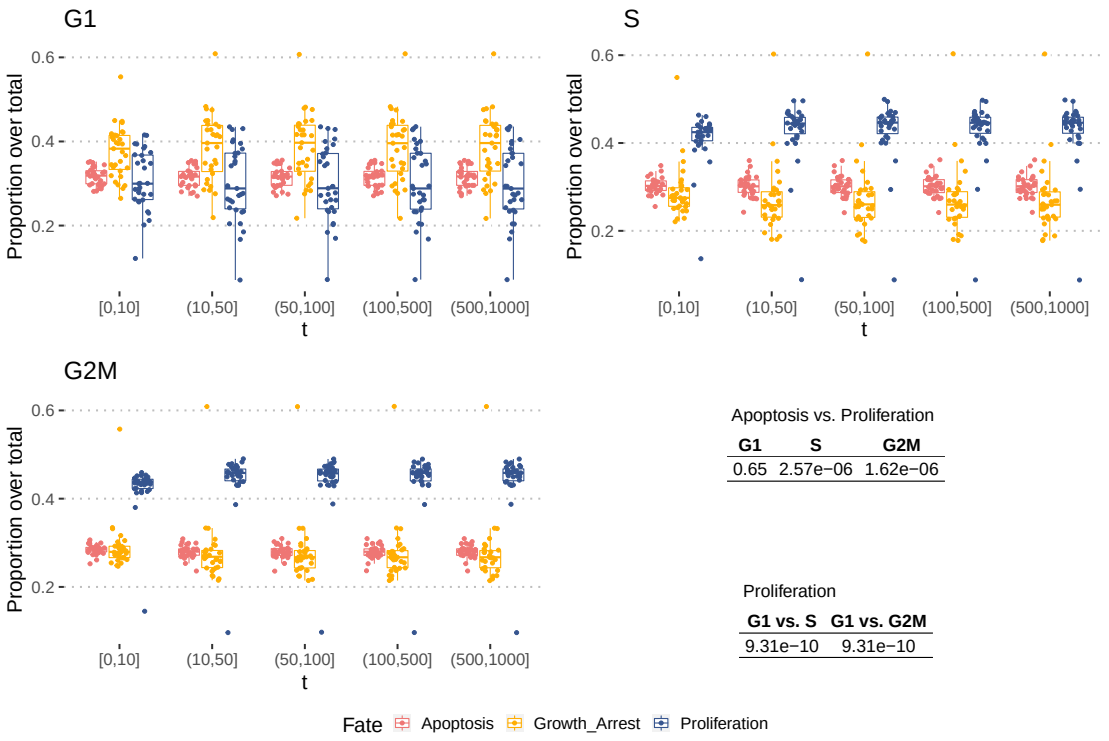


Figure 3.15: State transition computation for the networks produced by *Ingres*. Using every breast cancer cell line in the single-cell atlas as input, I computed 1000 state transitions for each of the three clusters networks generated by *Ingres*, G1, S and G2M. I ran this simulation on 1000 replicates for each network, due to the stochastic nature of PBN transitions. Then, I noted how many networks activated the fate nodes Proliferation, Apoptosis or Growth Arrest at each transition. The figure shows the proportion of each fate over the total at each time step (binned together for clarity, as the proportions barely change after 100 transitions), for each of the 32 cell lines. The tables show the p-values of pair-wise Wilcoxon tests comparing Apoptosis vs. Proliferation at each cell cycle phase, and the difference in Proliferation between the cell cycle phases, both using the 500-1000 time point data.

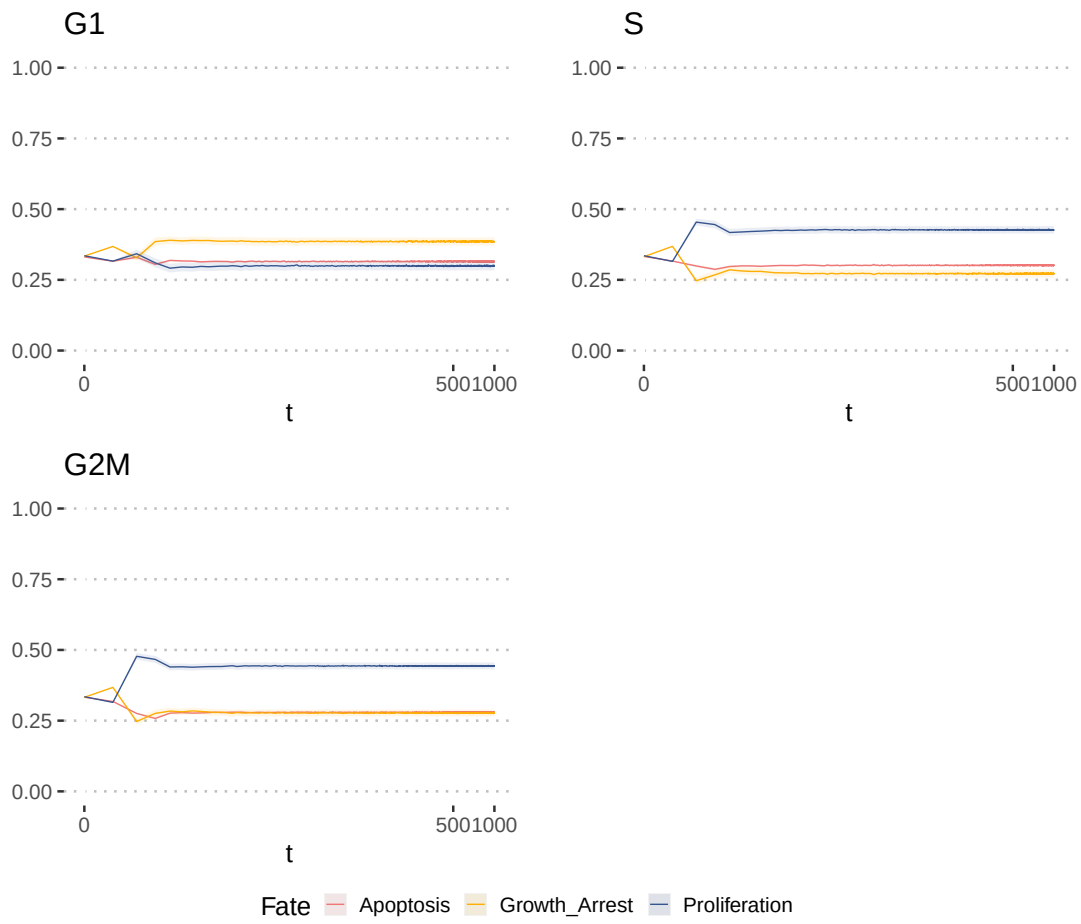


Figure 3.16: State transition computation for the networks produced by *Ingres*. The data used for this figure was produced as described in 3.15. The figure shows the proportion of each fate over the total at each state transition (shown with a pseudo-logarithmic scale), averaged over the 32 cell lines. The shading of the lines represents the standard error of the mean.

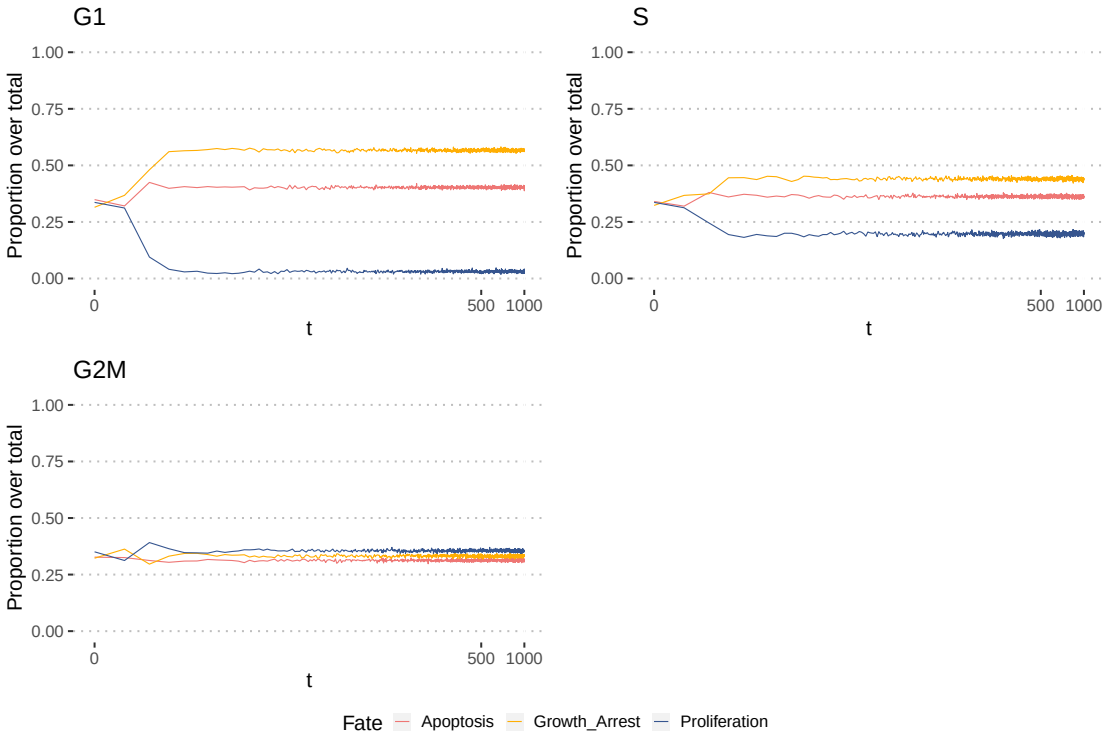
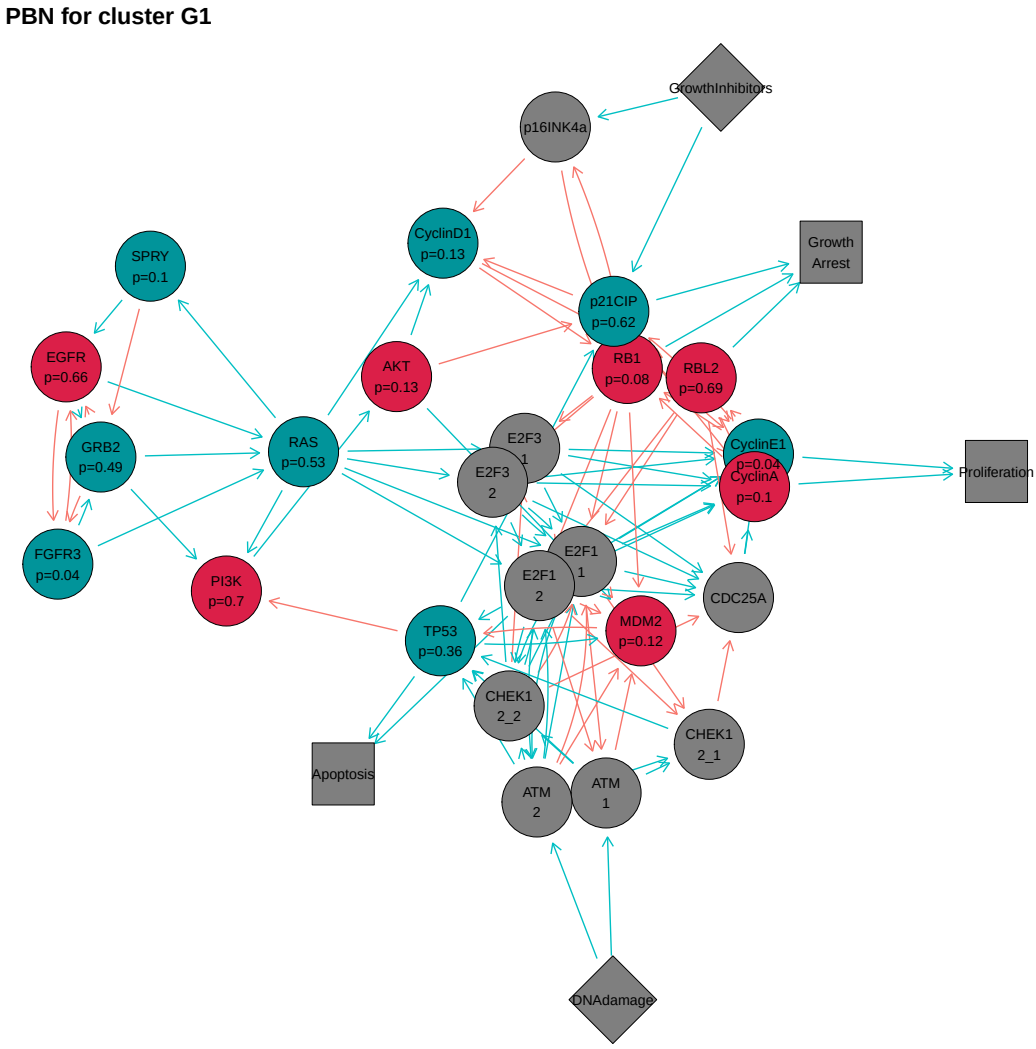
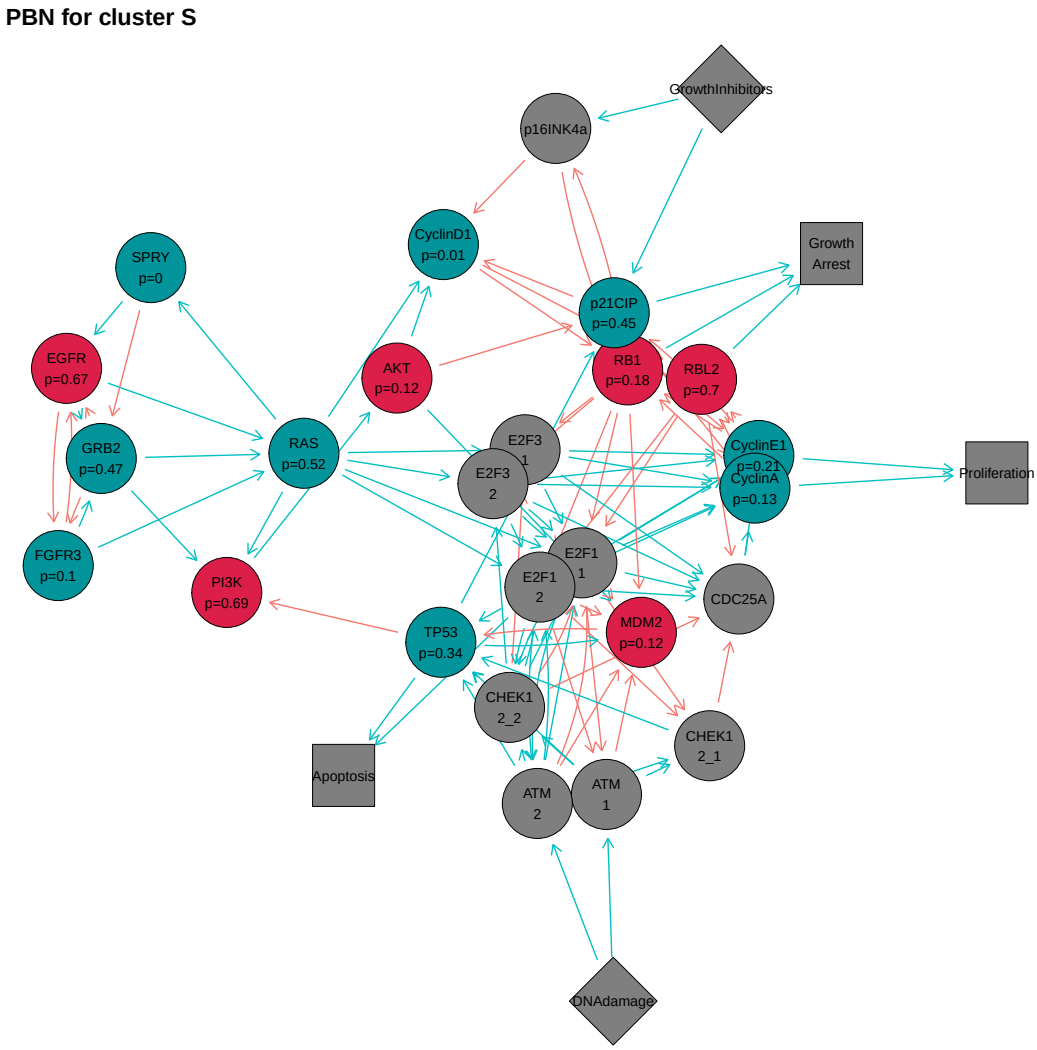


Figure 3.17: State transition computation for the networks produced by *Ingres* using a bladder cancer dataset. The procedure for this simulation was the same as described for Figure 3.15. The figure was produced as described for Figure 3.16.

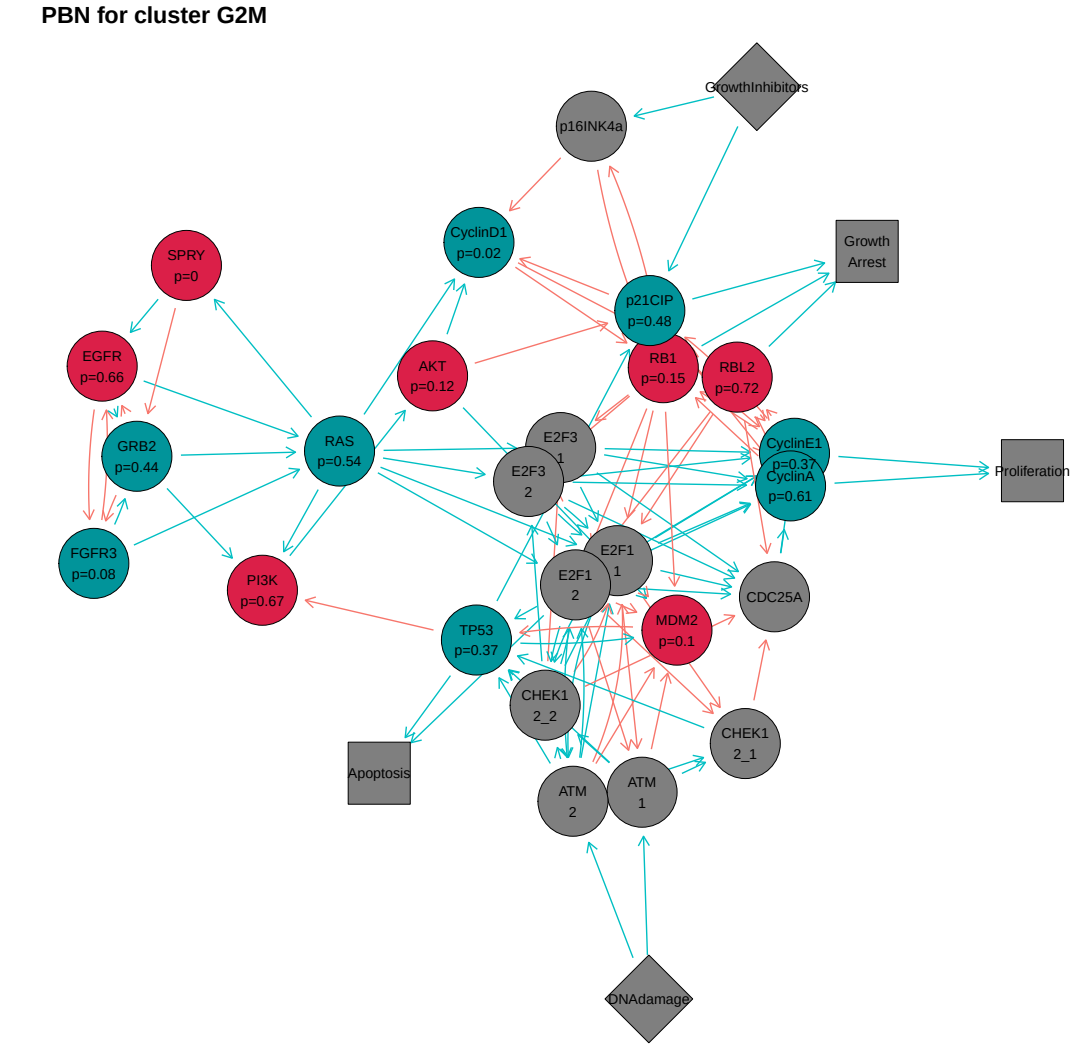
Figure 3.18: Networks produced by *Ingres* for each of the clusters using the bladder cancer dataset. Each of the PBNs produced by fitting the original Boolean network with the cluster data. ((c)): Summary of the networks probabilities as a heatmap. The nodes in grey in A-C and not present in D were not fitted, either because they are an abstraction of several nodes or because they were not found by VIPER.



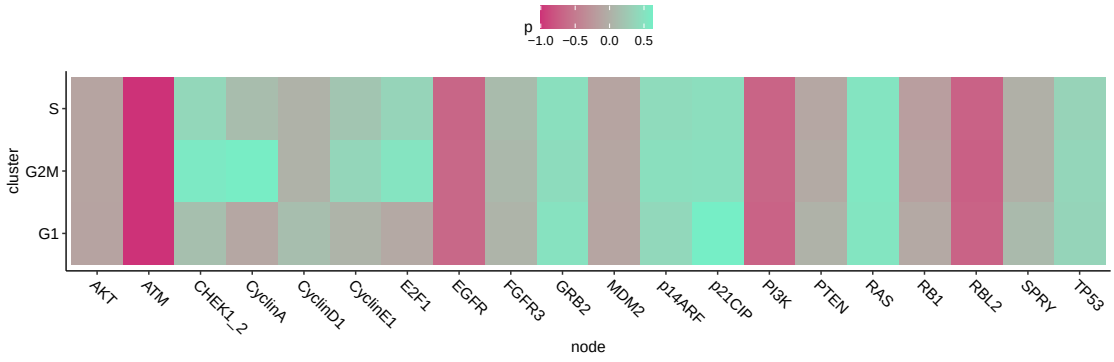
(a)


PBN for cluster S
GrowthInhibitors
p16INK4a
CyclinD1
p=0.01
SPRY
p=0
Growth
Arrest
p21CIP
p=0.45
EGFR
p=0.67
AKT
p=0.12
RB1
p=0.18
RBL2
p=0.7
GRB2
p=0.47
RAS
p=0.52
E2F3
1
E2F3
2
CyclinE1
p=0.21
CyclinA
p=0.13
Proliferation
FGFR3
p=0.1
PI3K
p=0.69
E2F1
1
E2F1
2
CDC25A
TP53
p=0.34
MDM2
p=0.12
CHEK1
2_2
CHEK1
2_1
Apoptosis
ATM
2
ATM
1
DNAdamage

(b)


PBN for cluster G2M
GrowthInhibitors
p16INK4a
CyclinD1
p=0.02
SPRY
p=0
Growth
Arrest
p21CIP
p=0.48
EGFR
p=0.66
AKT
p=0.12
RB1
p=0.15
RBL2
p=0.72
GRB2
p=0.44
RAS
p=0.54
E2F3
1
E2F3
2
CyclinE1
p=0.37
CyclinA
p=0.61
Proliferation
FGFR3
p=0.08
PI3K
p=0.67
E2F1
1
E2F1
2
CDC25A
MDM2
p=0.1
TP53
p=0.37
CHEK1
2_2
CHEK1
2_1
Apoptosis
ATM
2
ATM
1
DNAdamage

(c)


p
−1.0 −0.5 0.0 0.5
cluster
S
G2M
G1
AKT
ATM
CHEK1_2
CyclinA
CyclinD1
CyclinE1
E2F1
EGFR
FGFR3
GRB2
MDM2
p14ARF
p21CIP
PI3K
PTEN
RAS
RB1
RBL2
SPRY
TP53
node

(d)

3.3 Conclusions

In this chapter, I have:

- Tested the MAPK network both as an asynchronous and a synchronous network, and verified that I can obtain the same biologically relevant attractors the authors reported for asynchronous networks only. This is an requisite to ensure results by my model *microCJ* are reproducible regardless of the network update strategy used.
- Presented the pipeline *Ingres*, which merges a Boolean network with single-cell RNA-seq data, thus producing a probabilistic Boolean network.
- Verified that the PBNs produced by *Ingres* for different clusters in the same data-set are qualitatively different in their dynamic behaviour, and that this behaviour meets the expectations from the gene expression analysis of the cluster.

Ingres takes into account each network gene's whole regulon instead of just the gene's expression level, allowing to account for the global RNA-seq information into the final PBN, rather than just the expression of the single gene node. Using PBNs this way allows to model the difference of gene expression levels between cells or clusters, therefore allowing researchers to simulate scenarios that are more complex than total silencing or over-expression of genes.

I believe this approach is a good compromise between the simplicity and flexibility of Boolean networks and the wealth of information that single-cell sequencing allows. *Ingres* facilitates fitting models with cell-specific expression information without the need of inferring a new network for each cell or cluster, thus making the result more comparable from cell to cell, and more easy to scale in cell-based simulations involving hundreds of cells. Indeed, the resultant networks still have discrete states and outputs, which makes them perfect candidates to use in combination with agent-based modelling of cell and tissue dynamics. I expect it will prove useful for the biological modelling community at large. However, in the future it would be

interesting to combine this approach with single-cell protein levels using CITE-seq or similar technologies, to provide a more truthful representation of the proteomics landscape at the single-cell level.

3.4 Methods

3.4.1 Network analyses

I used *BoolNet*[316] version 2.1.7 along with R version 4.2.1 for the analyses in this section. The figures were produced with the package *ggplot2*[304]. The network was initialised with the inputs set to off:

```
network = fixGenes(network,
                   c("Oxygen_supply",
                     "EGFR_stimulus", "FGFR3_stimulus",
                     "TGFBR_stimulus", "DNA_damage", "
                     gefitinib_supply"),
                   c(1,0, 0,0,0,0))
```

Then, the attractors were found using one of the following three methods:

```
attr = getAttractors(network) # for synchronous networks
  attr = getAttractors(networktype='asynchronous',
  startStates = 1E7) # for asynchronous networks

attr = getAttractors(network,
  genesON = gene, startStates = 1E7) # for synchronous networks
    where a transition table is needed

transitions = getTransitionTable(attr)
```

When used with the MAPK network, due to its size, the first method performs by default the synchronous exhaustive SAT-based search as described in section 3.1.1. The second method performs a heuristic asynchronous search. Finally, the third one uses a heuristic synchronous state space search, which returns a transition table, unlike the first method. This was used for the fate activation-initial state plots, such as figure 3.5. Further details on these methods can be found on the *BoolNet*[316] documentation.

Regarding the box plots in the fate activation-initial state plots, with *ggplot2* the mid line is the median value, the lower and upper hinges correspond to the first and third quartiles, and the whiskers extend to the largest and smallest value,

respectively, no further than 1.5 * inter-quartile range. Data beyond the end of the whiskers are plotted individually.

3.4.2 Ingres

Ingres uses the latest available version of *VIPER*. *Ingres* source code is publicly available at `https://github.com/CBigOxf/ingres`. It includes documentation for every function and a tutorial vignette, included in this thesis as appendix A. Every function in *Ingres*, where possible, is subject to unit testing. The code conform to the guidelines from CRAN and Bioconductor, and is currently in submission to the latter.

For the Seurat analysis in section 3.2.2, I followed the Seurat vignette for cell-cycle scoring and regression[337]. The data was loaded using the function:

```
seurat.object = CreateSeuratObject(counts = data,
                    project = "ingres",
                    min.cells = 3,
                    min.features = 200)
```

and after that the executed code is identical to the aforementioned vignette. They were processed using the standard *Ingres* pipeline:

```
ingres = createIngresObjectFromSeurat(seurat.object, "RNA", "data",
          read_csv("networkGenes.csv")),
          graphmlAsTidy(paste0("network.graphml")))
          performViper(regulons) %>% # all regulons loaded from aracne.networks
          computePbnByCluster()
```

State transitions for figure 3.15 were computed using *BoolNet*. I computed 1000 state transitions for each of the three clusters networks generated by *Ingres*, G1, S and G2M. I ran state transitions on 1000 replicates for each network, due to the stochastic nature of PBN transitions. Then, I noted how many networks activated the fate nodes Proliferation, Apoptosis or Growth Arrest at each transition. The proportions of each fate activation over the total was computed. I performed pair-wise Wilcoxon tests comparing Apoptosis vs. Proliferation at each cell cycle phase, and the difference in Proliferation between the cell cycle phases, both using the 500-1000 time point data. Results were plotted using *ggplot2*[304].

All models are wrong, but some are useful.

— George Box

4

Description of the computational tumour model

Contents

4.1 Features 124
4.1.1 Network modelling 124
4.1.1.1 Fate decision 126
4.1.1.2 Import transcriptional data 127
4.1.1.3 Random mutations 129
4.1.2 Cell placement 130
4.1.3 Diffusion 131
4.2 Results 134
4.2.1 Effect of mutations on the model behaviour 134
4.2.2 Competitive advantage in multiclonal simulations 137
4.2.3 The role of the microenvironment in the model evolution 141
4.2.4 Probabilistic Boolean Networks for a more granular simulation 150
4.2.5 Acquired mutations and clonal evolution 152
4.2.6 Drug treatment 162
4.2.7 Sensitivity analyses 168
4.2.7.1 Instant growth arrest 168
4.2.7.2 Decision window 168
4.2.7.3 Growth arrest duration 172
4.2.7.4 Shoving 173
4.2.7.5 Cell placement method 174
4.3 Conclusions 183
4.4 Methods 185
4.4.1 Implementation 185
4.4.2 CLI and GUI 188

4.4.3 Settings 189

In this chapter I will describe *microCJ*, a computational multi-scale framework for cancer modelling (figure 4.1) the development of which has been the main focus of my DPhil project.

microCJ is inspired by microC[257], a agent based modelling framework developed in my lab. microC merges several levels of simulation: intracellular interactions through a Boolean network, intercellular interactions through a on-lattice model of cell growth, and interaction with the microenvironment through a diffusion model. *microCJ* starts from this concept and takes it much further: it adds many new features to the intracellular level, such as the ability to simulate asynchronous and probabilistic Boolean networks, probabilistic mutations and cells acquiring new mutations during the simulation; to the intercellular level such as different algorithms for the placement of cells both on-lattice and off-lattice or the capacity of cells to push others to make space for new growth; and to the diffusion level such as temporary drug administration schedules. This is implemented as an open source modular Java program, in such a way that is easy to extend and modify, which I believe were shortcomings of microC. Moreover, *microCJ* is built with the integration with *Ingres*, the other computational tool described in this thesis (chapter 3). It will easily import single-cell RNAseq data as a virtual cell per each real cell, controlled by a PBN derived from the gene expression data for that specific cell. All in all, I believe *microCJ* to be a sophisticated modelling framework that brings together a lot of the current ideas in the field, such as the integration of transcriptomics data and network models, or off-lattice 3D growth models, into a cohesive and modular code base.

As previously mentioned, *microCJ* follows the agent-based paradigm. Models of this kind are composed of autonomous, self contained entities, called agents, which interact with other agents and the environment, changing their behaviour of properties as a result. The main advantage of agent-based models is that they are able to represent heterogeneous and complex systems. Each agent has the same set

of actions available, but each of them have different properties that will influence which action is executed and how that action is specifically implemented. For example, in *microCJ*, all cells have the ability to divide, but whether this happens or not for a given cell, or the rate at which a cell divides, will be determined by that cell's specific properties: its age, its mutations, whether there is enough space in the environment, and so on.

In *microCJ*, agent based modelling actually happens on three levels. First, as illustrated in figure 4.1, each cell is an agent, which a series of actions available to it, mainly: to divide, to undergo apoptosis, to consume oxygen and to produce growth factors; and a series of properties that may vary for each cell: its age, its location, its oxygen consumption rate, the activation status of each of its genes and its mutations.

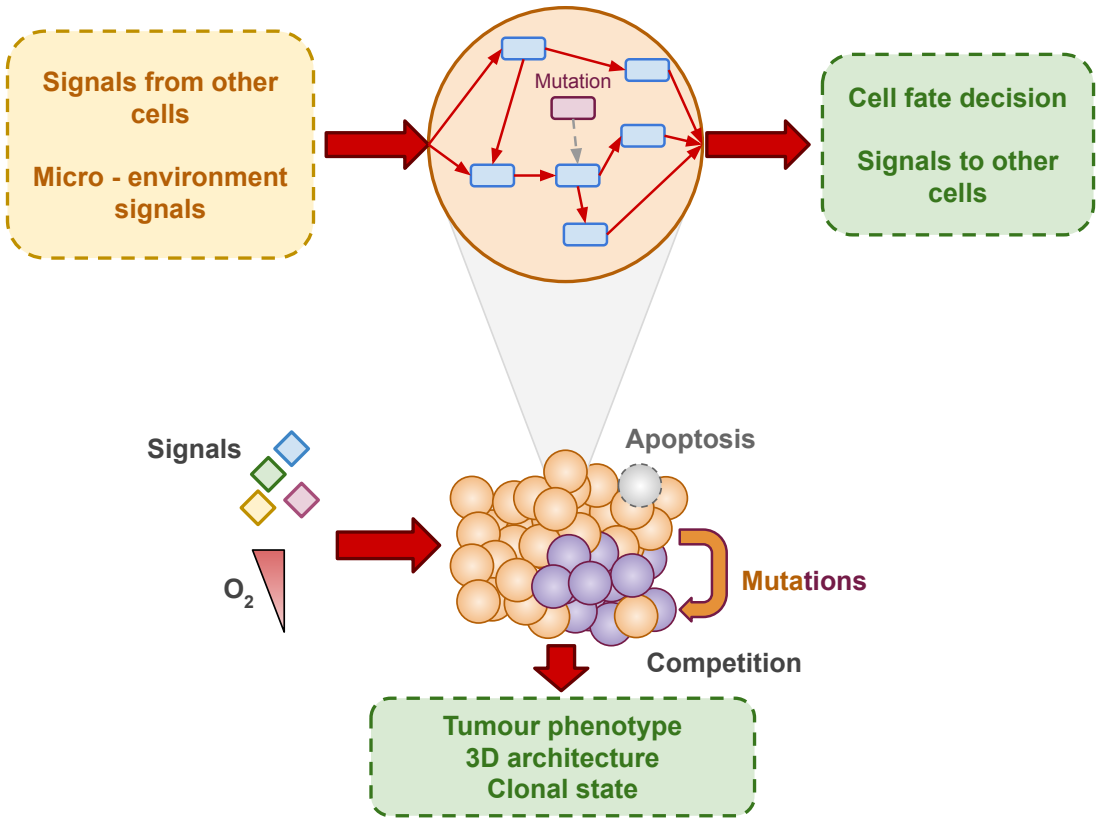


Figure 4.1: Overview of *microCJ* functionality.

Secondly, for each cell there is a second layer of modelling, where a genetic regulatory network (GRN) is simulated. This kind of network is a high-level

abstraction of regulatory networks that include genetic expression, RNA regulation and protein interaction. Nodes are connected through directed links that represent their interaction, regulatory activity or feedback loop. *microCJ* accepts synchronous, asynchronous and probabilistic Boolean networks (PBNs, chapter 3, page 89). The output of the network is a cell fate decision. Fate decisions are represented as end nodes, meaning nodes with incoming links but no outgoing links. Common examples of such nodes are proliferation, apoptosis, necrosis, hypoxia or growth arrest. When the network reaches a stable state in which one of the fate nodes is on, that fate is implemented by the cell to which that network belongs to.

Any gene on any given cell can be mutated, the exact meaning of which will depend of the kind of network used in the simulation. For synchronous and asynchronous Boolean networks, it means that the node will be fixed to one state, on or off. For PBNs, it means that the probability of each rule can be modified, or new rules added. This allows to model tumour heterogeneity and clonal subpopulations, since each cell can have a different set of mutations, and mutations are passed down to daughter cells.

Lastly, oxygen, growth factors and other substances' diffusion across the tumour can also be simulated. When concentration of oxygen or a growth factor around a cell reaches a certain threshold, the corresponding node in the network of that cell becomes activated, and the concentration of that molecule in the vicinity of the cell is decreased.

In the same way, when certain nodes of the network are activated, the cell will produce a growth factor that will diffuse and may be consumed by neighbouring cells.

4.1 Features

4.1.1 Network modelling

At the core of the *microCJ* model is the network simulation. For every cell simulated in the model, there is a corresponding network that governs the behaviour of the cell. The network inside each cell run independently from other cells' networks,

but cells can interact with one another via growth factors, competition for nutrients and space, as will be discussed later.

The networks used are Boolean gene regulatory networks (BGRN), as described in chapter 3. There are three types of nodes in the networks used by *microCJ*. First, there are nodes which have both incoming and outgoing edges, and represent the genes or proteins involved in a specific pathway. Second, there are input nodes, which only have outgoing edges. Their state is not set by a Boolean function, but it depends on factors external to the network. They represent the agents foreign to the specific process that it is being modelled, such as a receptor protein becoming activated, an immune response being triggered, or a nutrient being absorbed by the cell. Finally, there are fate nodes, which only have incoming edges. Their state is not read by other nodes in the network, they represent the phenotype of the network, for example, they might represent the proliferation or apoptosis of the cell to which the network belongs.

Furthermore, cells in the model can present over-expression, under-expression or deletion mutations, via changing the behaviour of specific nodes in their networks. This allows to model different sub-populations within the model, with different genotypes. *microCJ* also includes a system through which users can define mutations that appear at specific time points for a fraction of the population, or even the possibility for cells to autonomously gain mutations after division. Thus, the effect of the order in which mutations appear or of intratumour selection have in the overall phenotype of the tumour can be simulated.

The exact way in which these mutations are modelled depends on the type of network used for the simulation — *microCJ* can work with the three different kinds of BGRN: synchronous, asynchronous and probabilistic.

When synchronous or asynchronous networks are used with *microCJ*, mutations are modelled as nodes being fixed to one or zero, representing over-expression or deletion, respectively. That is, when the mutated node is to be updated, it will stay fixed to one or zero, regardless of its Boolean rule or the state of other

nodes. This has the limitation that different levels of expression between subpopulations can't be modelled.

Mutations in PBNs are represented by adding Boolean functions that always equal one or zero, but with varying probability, as to model differential expression between cells. For example, a node A_1 can have the rules [!B & C] p = 0.8*, [1] p = 0.2, and a node A_2 can have the rules [!B & C] p = 0.6, [1] p = 0.4. In this example, both A_1 and A_2 will follow their first rule most of the time, but A_1 will have a bias twice as big as A_2 towards activation. This way more granular representation of differential expression can be achieved but still confined to Boolean networks, without the added complexity or need for extensive knowledge of quantitative models.

microCJ includes an integration to import the networks produced by *Ingres* and load them into the model, so that each cell in the *microCJ* simulation correspond to an actual cell in the single-cell RNAseq dataset, its behaviour determined by a PBN fitted with the real cell gene expression matrix.

4.1.1.1 Fate decision

The fate decision connect the intracellular level of the model to the intercellular one, by causing cells to divide, to die or to stop growing. There are four main possible fates:

- Proliferation: the cell creates a new daughter cell in an adjacent location in space.
- Growth arrest: the cell stops updating its network and its consumption and production of diffusible substances halves.
- Necrosis: caused by lack of oxygen intake. The cell dies and stops updating its network and state, it does not consume or produce any substances, but keeps on occupying space.

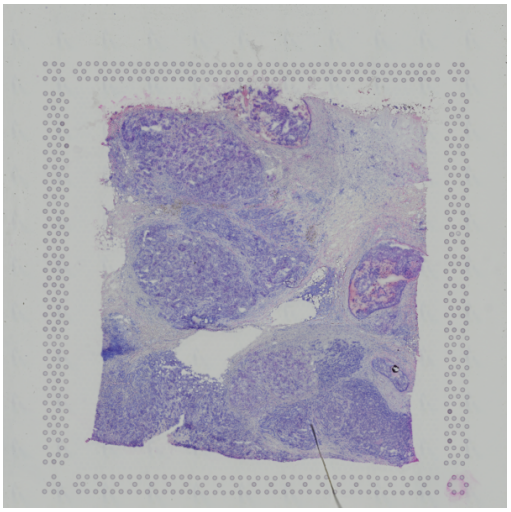
*Boolean rule notation meaning that the node in question will be active if and only if node B is inactive and node C is active, and otherwise it will be inactive. p denotes the probability that this rule is chosen at a given time step.

- Apoptosis: the cell dies and it is removed from the simulation, leaving an empty space behind.

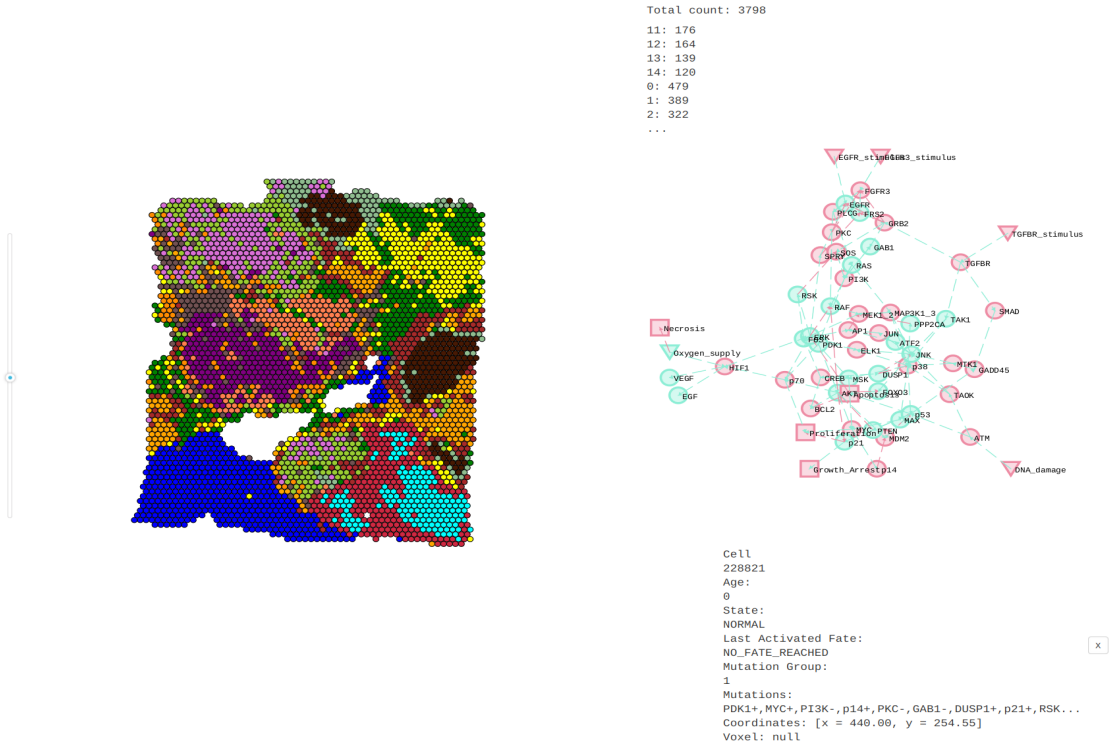
The effects of necrosis and apoptosis are executed immediately, this is, in the time step where the corresponding node becomes activated, the cells change their state and are removed, in the case of apoptosis, before the start of the next time step. On the other hand, proliferation and growth arrest are only executed at the end of a so-called decision cycle. The length of this cycle is named the *decision window*, and users of the model can change this in the settings. Every cell starts the simulation with a random position inside the decision window, that is, some may end their cycle sooner than others. This is to prevent different cells, maybe from different clonal subpopulations, to advance through the cell cycle and divide in unison. At the end of the decision cycle, the cell may go into proliferation or growth arrest as soon as the corresponding node is activated, after which the related effects will be executed and the decision window will restart. This is meant to model the duration of the cell cycle and the need for the cell to grow in between divisions. As growth arrest is a possible outcome of the cell cycle, it is modelled in the same system as proliferation, but this can be deactivated in the model settings, so that growth arrest happens immediately after node activation. The amount of time a cell remains arrested before turning to its normal state again is also a configurable variable.

4.1.1.2 Import transcriptional data

Using *Ingres* (chapter 3, page 105), a single-cell or spatial RNA-seq dataset can be loaded into *microCJ*. In both cases, *microCJ* accepts either the cluster networks or the single-cell networks produced by *Ingres*.



(a)



(b)

Figure 4.2: *microCJ* can import spatial RNA-seq data. The data for this example comes from the publicly available Visium breast cancer dataset. (a) shows the image of the tissue sample and (b) shows the *microCJ* graphical interface and the cells placed in the same position as the 'beads' in the Visium dataset, coloured by cluster.

If cluster networks are used, the user can specify the total number of cells in the simulation and each cluster will have the same relative size (in number of cells) as in the original dataset. If the single-cell networks are loaded, then a *microCJ* cell is created for each actual cell in the dataset. In the case of spatial data, the same is true, and furthermore, the spatial information is also loaded, so every cell in *microCJ* will be placed in the same relative position as in the dataset (figure 4.2). *Ingres* networks are loaded as the specific PBNs for every cell, each of them with the original rule from the synchronous BN for every node, and the additional rule for a fix to one or zero, with a probability relative to its protein enrichment score. Therefore, every cell in *microCJ* will be running a different network. When one of the cells proliferates, the daughter will inherit that cell's network.

4.1.1.3 Random mutations

The user can define mutations and group them into different subpopulations for them to be used throughout the whole simulation. But the model can also be configured so that cells can acquire new mutations every time they divide. An *acquired mutation probability* parameter can be set, and every time a new cell appear, it will inherit its parent's network, but it will also add a mutation to any of its gene nodes with that given probability. When a new mutation is acquired, its orientation (if it is a fixed one or fixed zero mutation) is decided at random. For probabilistic cells, both the orientation and its probability are decided at random. A node can accumulate several mutations through the course of the simulation, with certain limitations. In a synchronous network, deletion mutations (fixed zero) cannot mutate directly back to wild type or over-expression, but they can go for over-expression to wild type and deletion, but it will take two sequential mutations (deletion to wild type, wild type to over-expression). This can happen in biology, for example, tumours exhibiting HRD can restore HRR by inactivation of 53PB1 or other proteins, but also through secondary mutations in BRCA1/2 that restore the open reading frame of the mutant BRCA allele and therefore its function[60]. For

probabilistic networks, the probability that an already mutated node mutates again is inversely proportional to the mutation's current value, as per the formula:

$$p_{total} = \frac{p_{acquisition}}{c_{node}}$$

where p_{total} is the total probability that the node will mutate in a given time step, $p_{acquisition}$ the base acquired mutation probability and c_{node} the existing probability for the already existing mutation, if any, in the node.

4.1.2 Cell placement

Cells occupy physical space in the simulation, that can be run in two or three dimensions. Cells are abstracted as circles or spheres of a given radius (r). *microCJ* offers different option for the placement of cells, according to several methods of circle/sphere packing. For 2D simulations, the cells can be placed in a square grid — every cell has eight neighbour cells —, a hexagonal grid — six neighbours — or at random, where every new cell will be placed in a random point at a distance of $2r$ from the center of the existing cell. For 3D simulations, the options are similar: square grid — 26 neighbours — and random placement.

When a cell activates its Proliferation fate, it then checks if there is such a point at a distance of $2r$ where a whole cell could be placed, following the placement option chosen, without overlapping with any existing cell. If this check is passed, the cell divides normally, which in practical effects means that a new cell is created, preserving the mutations of the parent cells, if any, and placed at that point. The user can define which of two options are executed if the check is not passed. First, that cell simply doesn't divide, and instead goes into growth arrest for a given period of time. Second, the cell pushes all the neighbouring cells out by a small distance, checks if there is now enough space for the new cell, and repeats until there is. This shoving algorithm is executed by all overlapping cells until an arrangement where no cells are overlapping is reached.

4.1.3 Diffusion

The space in the simulation is divided in a 2D or 3D grid for the purpose of diffusion simulation. The placement of cells described in the previous section is independent from this grid. Cells are assumed to be in only one square or voxel of the grid, as determined by the position of their geometrical center. Each voxel has a specific concentration of each of the diffusible substances defined by the user. The initial concentration of each substance can be modified in the settings before the start of the simulation, as well as the diffusion coefficient and the consumption and production (if applicable) rate for each substance.

The diffusion of substances in *microCJ* is simulated following the diffusion-reaction equation:

$$\frac{\partial S}{\partial t} = D_s \nabla^2 S + R_{S\rho_{cell,s}} \tag{4.1}$$

where S is the concentration for substance s, D_s its diffusion coefficient, and R_s its rate of consumption and production. $\rho_{cell,s}$ represents the density of cells which consume or produce substance s.

microCJ uses the Forward Time Central Space (FTCS) method to discretise the diffusion-reaction equation at the centre of every voxel:

$$\frac{\partial S}{\partial t} = D_s \left[\frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2} + \frac{\partial^2 S}{\partial z^2} \right] + R_s \tag{4.2}$$

We can refer to the concentration of a given substance (such as oxygen or glucose) at a voxel as *S(i,j,k)*, where *i, j, k* denotes the coordinates of the voxel. Then, knowing the concentration of a substance in every voxel in the simulation at $t^n = n \cdot dt$, expressed as $S^n_{i,j,k}$, the concentration of that substance at $t^{n+1} = (n+1) \cdot dt$ can be calculated for every voxel using the following equation:

$$S^{n+1}_{i,j,k} = S^n_{i,j,k} + dt \cdot D \left[\frac{S^n_{i+1,j,k} - 2S^n_{i,j,k} + S^n_{i-1,j,k}}{dx^2} + \frac{S^n_{i,j+1,k} - 2S^n_{i,j,k} + S^n_{i-1,j-1,k}}{dy^2} + \frac{S^n_{i,j,k+1} - 2S^n_{i,j,k} + S^n_{i-1,j,k-1}}{dz^2} \right] + dt \cdot R^n_{i,j,k} \tag{4.3}$$

Where $R^n_{i,j,k}$ denotes the substance consumption and production (if applicable) of all cells residing in voxel *i,j,k*.

$$R^n_{i,j,k} = C^n_{i,j,k} + P^n_{i,j,k} \tag{4.4}$$

$C^n_{i,j,k}$ is the total consumption of cells in that voxel, K_c is the consumption rate and N_c is the number of alive cells in the voxel.

$$C^n_{i,j,k} = -K_p \cdot N_c \cdot S^n_{i,j,k} \tag{4.5}$$

$P^n_{i,j,k}$ is the total Substance production of cells in that voxel, K_p is the production rate and N_p is the number of producing cells in the voxel.

$$P^n_{i,j,k} = K_p \cdot N_p \tag{4.6}$$

Oxygen consumption follows a special equation:

$$C_{O_2} = -K_{O_2,p} \frac{S_{O_2} - S_{NS}}{S_{OP} - S_{NS}} N_{O_2,c} \tag{4.7}$$

Where S_{NS} is the lowest possible oxygen concentration (normally 80% of the oxygen activation threshold) and S_{OP} is the optimal oxygen concentration, currently set to 0.28 mM, which is the high value I use for normoxic simulations. This was inspired by microC[257] , and it is meant to realistically model the way cells consume oxygen depending of its concentration (whether this is normoxic, hypoxic, or anoxic) compared to normoxic conditions.

microCJ sets dt to be:

$$dt = 0.1 \frac{dx^2}{D} \tag{4.8}$$

for the substance having the biggest D, oxygen in this case. For any other substance with a smaller D' value (slower diffusion, such as a growth factor):

$$dt = 0.1 \frac{dx^2}{D} \frac{D'}{dx^2} \tag{4.9}$$

For the simulation to be numerically stable, dt should satisfy:

$$dt < \frac{dx^2}{6D} \tag{4.10}$$

which is guaranteed by equation (4.8) and equation (4.9).

Typically, the time scale of substance diffusion is much smaller than that of cell proliferation (seconds versus hours) and therefore the diffusion is assumed to reach a steady state before the cell can divide. With the above equations and the typical parameters for *microCJ* of a voxel size of 20 µm and oxygen diffusion coefficient of 10^{-9} m^2/s, the maximum possible diffusion time step is 0.066 seconds (from equation (4.8)). If the simulation domain size is 200 µm each side, then the time a substance takes to diffuse from the boundary to the center is about 6.66 seconds (0.002 hours). This is much smaller than the usual cell division time of 24 hours. Therefore, the diffusion process can be considered to be in a quasi steady state compared to the cell division process. As such, it is needed to solve the diffusion equation to a steady state before cell growth. The number of diffusion steps it takes to diffuse oxygen from the boundary to the center is around 0.066/6.66 = 100 steps. This means the diffusion will arrive to the quasi steady state in around 100 iterations. Since the decision window for division I use for microCJ is usually 100-200 steps, the diffusion steady state will be reached before any division step, and therefore the two are effectively decoupled. Nevertheless, this is an area for improvement in *microCJ*, since the ideal modelling strategy would be to have different time scales for each level of the model.

At every diffusion step the cells consume substances at a rate determined by equation (4.5) and equation (4.7). If the cell intake goes over the activation threshold (as defined in the settings) for that substance, the corresponding node in the cell network gets activated. Otherwise, it is set to inactive.

Conversely, some output nodes in the regulatory network denote the production of substances by cells for other cells to consume, such as growth factors. When such a node gets activated in a cell's network, that cell counts as producing, and it will be counted in the N_p term in equation (4.6).

Boundary voxels maintain a constant concentration for all substances equal to their initial concentration (Dirichlet boundary conditions), so equation (4.3) does not apply to them. This models the nutrients being accessible to the tumour from the blood supply and diffusing inwards.

Drugs are a special case. They are simulated as a diffusible substance, but the initial concentration is not maintained constant at the boundary. Instead, the user can set the length of the treatment, during which the drug concentration at the boundary will be constant at a given value. Before and after treatment, concentration at the boundary will be 0.

4.2 Results

Throughout this section I will test every feature of the model to assess its effect on the general behaviour, as well as doing sensitivity analysis of the mail parameters. Unless otherwise stated, simulations started with 100 cells and were run for 100 replicates, using a synchronous network, in a 3-dimensional space. Cells were distributed randomly — in a non-lattice formation —. "Mutations" are meant to mean that a gene is fixed to one (+) or to zero (-) in a cell's network. The initial state of the network inside each cell is also random, as is its progress through the decision window. Appendix B contains all the parameters for every simulation included in this chapter.

4.2.1 Effect of mutations on the model behaviour

First, I evaluated the effect of certain mutations on the dynamic behaviour of the plot, decided according to the analyses presented on chapter 3 (page 89). Figure 4.3 shows the result of a simulation with populations EGFR+, EGFR+p53-, EGFR+p53-PTEN-, p53- and WT. It can be seen, as also shown on chapter 3 that combining EGFR over-expression with p53 deletion ensures high activation of the proliferation fate node, granting exponential growth to that population in *microCJ*. On the other hand, applying just one of those mutations on its own does not have

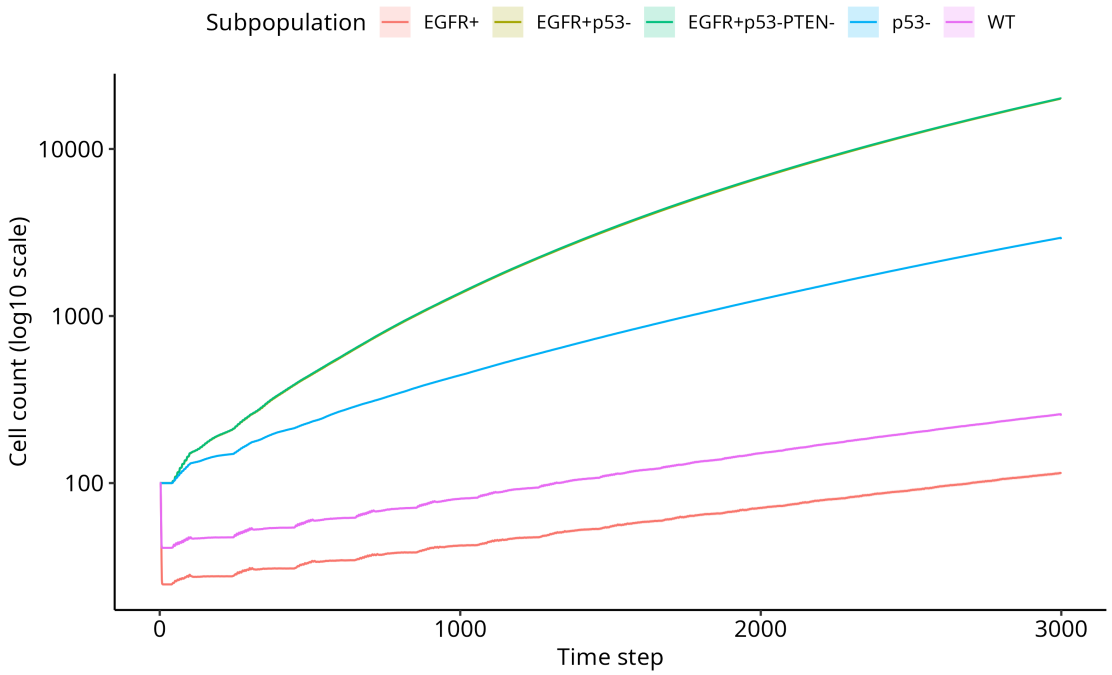


Figure 4.3: Populations with different mutations growing independently. Five populations are simulated, with different mutation sets. A '+' or a '-' symbol denotes that the gene has been fixed to one or to zero, respectively, during the whole simulation. The mutation sets are EGFR+, EGFR+p53-, EGFR+p53-PTEN-, p53- and WT (wild-type, no mutations). The figure shows the growth of every population across 3000 time steps. EGFR+p53-, although not visible in the plot, is behind EGFR+p53-PTEN-, their curves are almos identical. The y axis is the number of alive cells (i.e. not apoptotic nor necrotic) at each time step. Every population was simulated independently, with 100 replicates each. Each simulation started with 100 cells. The lines show the mean of the 100 replicates at each time point, and the shaded area the standard error of them, which in this plot is very narrow. EGF and oxygen diffusion were simulated too, with an oxygen boundary concentration of 0.28 mM (normoxia).

such a strong effect, and it is barely distinguishable from the wild-type population (WT, containing an unperturbed network) in the case of the EGFR+ population.

Taking a closer look at the expression profile of each population (figure 4.4) it can be seen that although all cells start with a random gene activation profile, by time point 200 they have reached the state that they will keep for the rest of the simulation, including any new cell born after the start. These states match the attractors for these perturbations shown on chapter 3, which is an interesting insight into how the cells, despite starting from a random initial state in their networks (that could be interpreted as a random point in the internal signalling dynamics) do

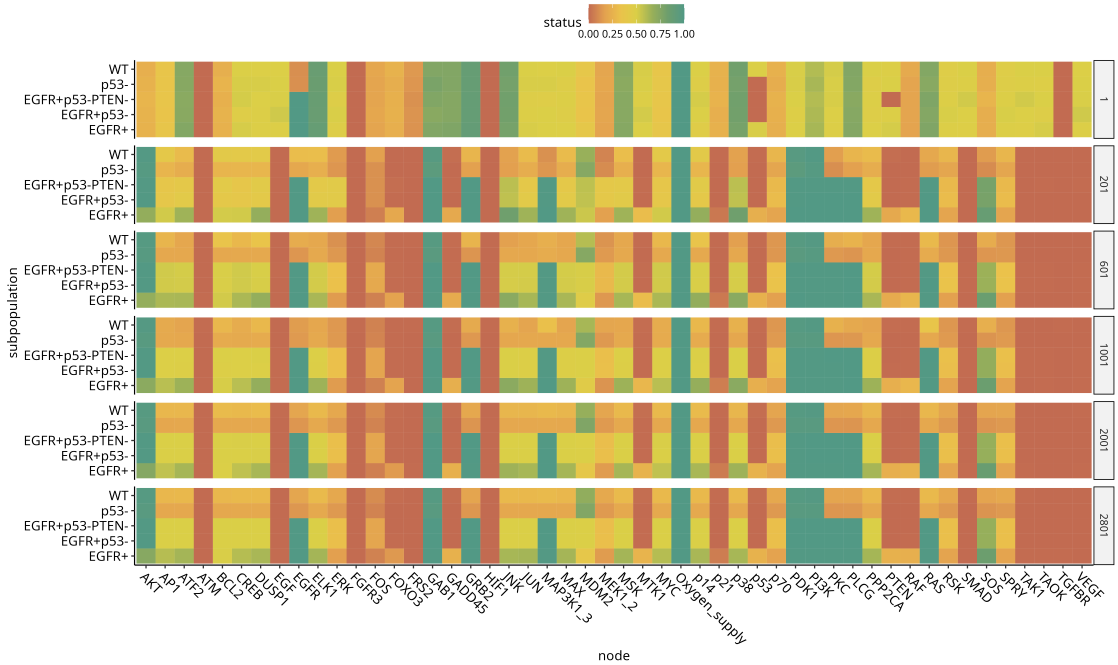


Figure 4.4: Average node activation of independent populations across different time points. Using data from the same simulation as figure 4.3, this figure shows the average status per network node across all alive cells in a population, and across a 100 replicates. The heatmaps show the results at several time points of the simulation: 1, 201, 601, 1001, 2001 and 2801. Each block of rows is a different time point, each row inside the block corresponds to a different subpopulation, as labelled on the left hand side vertical axis. These populations are: WT (wild type), p53-, EGFR+ p53- PTEN-, EGFR+ p53-, EGFR+.

not take very long to fall into the attractor corresponding to the perturbations in their network, or mutations, which represents the connection between genotype and phenotype. Because the cells are all normoxic and genetically homogeneous, their network is the main (and possibly only) factor in their growth. The populations that present a higher predicted proliferation (the ones with EGFR+ and/or p53-) have a higher average expression of proliferation-related genes such as ERK, RAS or MYC.

One can see that even the WT cells display activation of proliferation-related genes such as PI3K or AKT, and low expression of tumour suppresing genes such as PTEN or p53. But one have to bear in mind that most of the WT cells go into apoptosis or growth arrest shortly after the start of the simulation, and it is only the few cells that start with one of the network initial states that lead to proliferation

(as previously indicated, network initial status is set at random for every new cell in the system) that are predicted to survive, and thus the only ones of which the node status is read for the purposes of figures such as figure 4.4.

4.2.2 Competitive advantage in multiclonal simulations

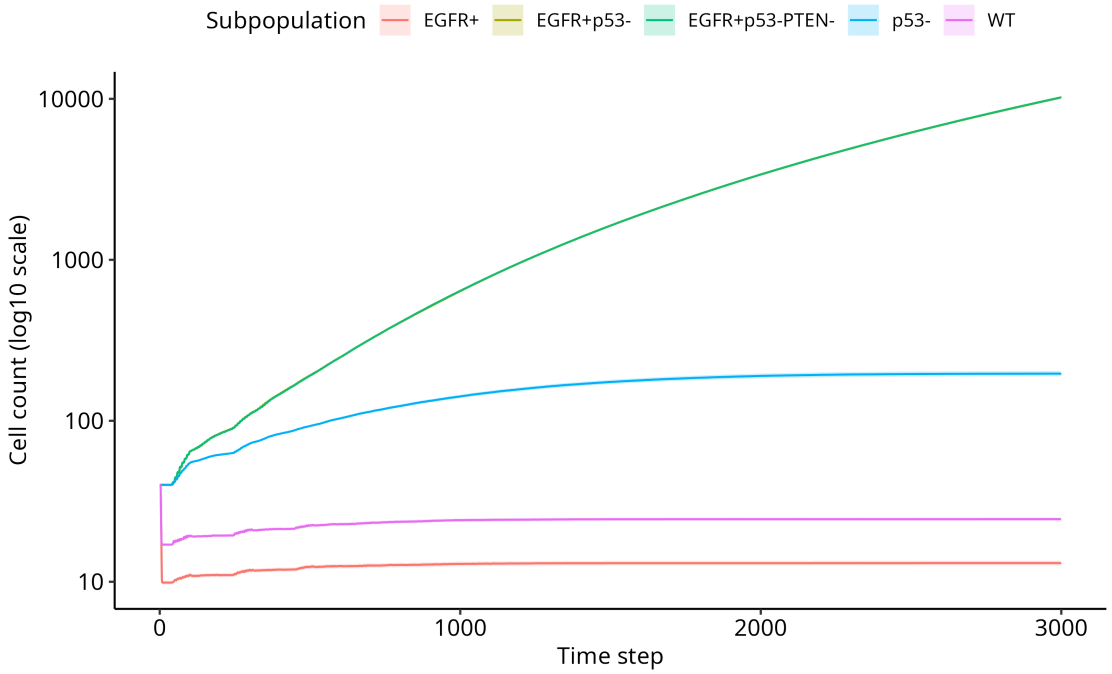


Figure 4.5: Effect of mutations on competitive performance. This simulation was run with the same parameters as figure 4.3, but instead of running a separate simulation for each population, all of them were included in the same one. Each of them had the same number of initial cells, 40 each.

Another matter to look at is how these mutations are predicted to confer an advantage or disadvantage when several subpopulations are competing inside a tumour. In figure 4.5 one can see the result of this. While the EGFR+ and p53- subpopulations remain the more proliferative by a large margin, the absolute number they reach after 3000 iterations is much lower, around the 10,000s rather than the 20,000s as in figure 4.3. However, the other competing subpopulations do much worse than in isolation, barely growing at all.

Their expression profile (figure 4.6) is almost identical to that of the populations growing in isolation (figure 4.4).

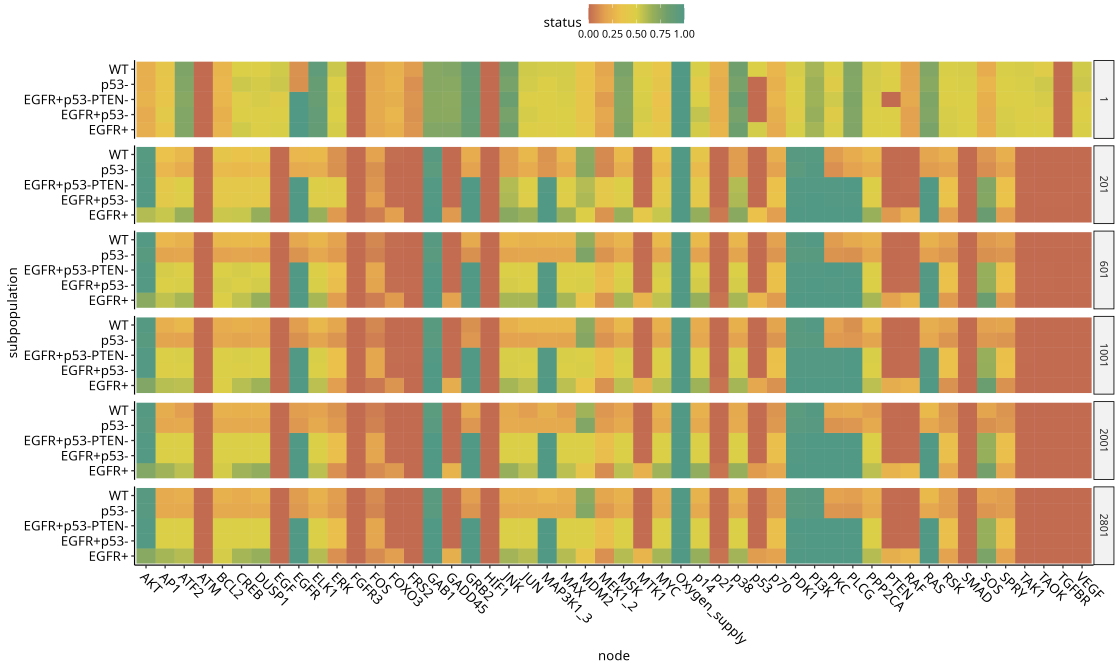


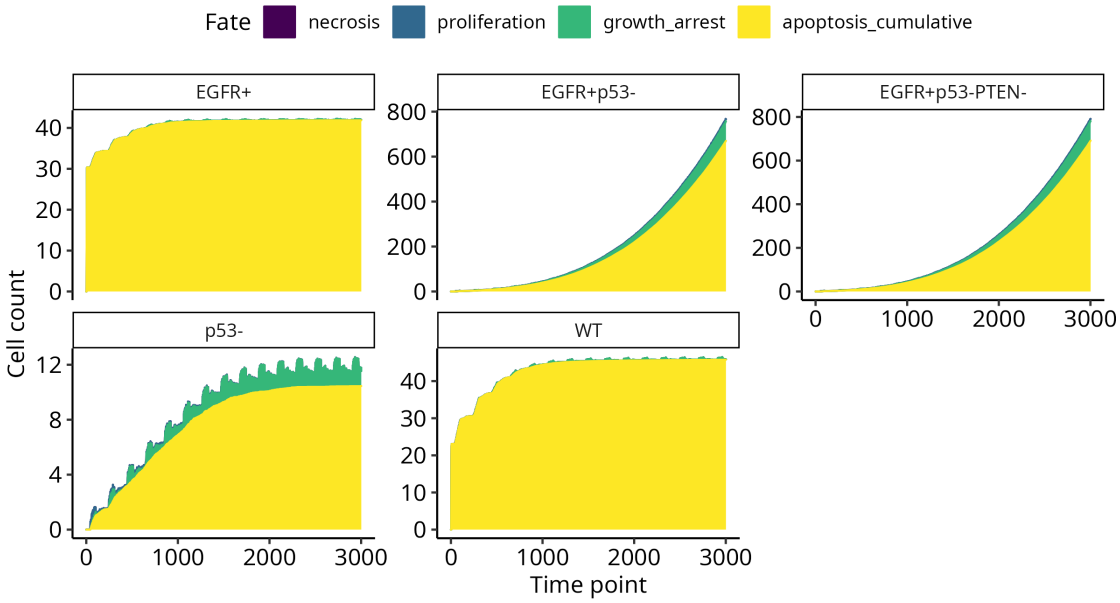
Figure 4.6: Average node activation of competing subpopulations across different time points. Data from the same simulation as figure 4.5. Figure legend as in figure 4.4

Looking at figure 4.7 one can see that the reason for this under-performance is that the most aggressive clones eventually take over the space and leave no room for the other subpopulations to grow, condemning them to growth arrest —as the model is configured, a cell that tries to proliferate but has no room to create a daughter cell, enters in growth arrest.

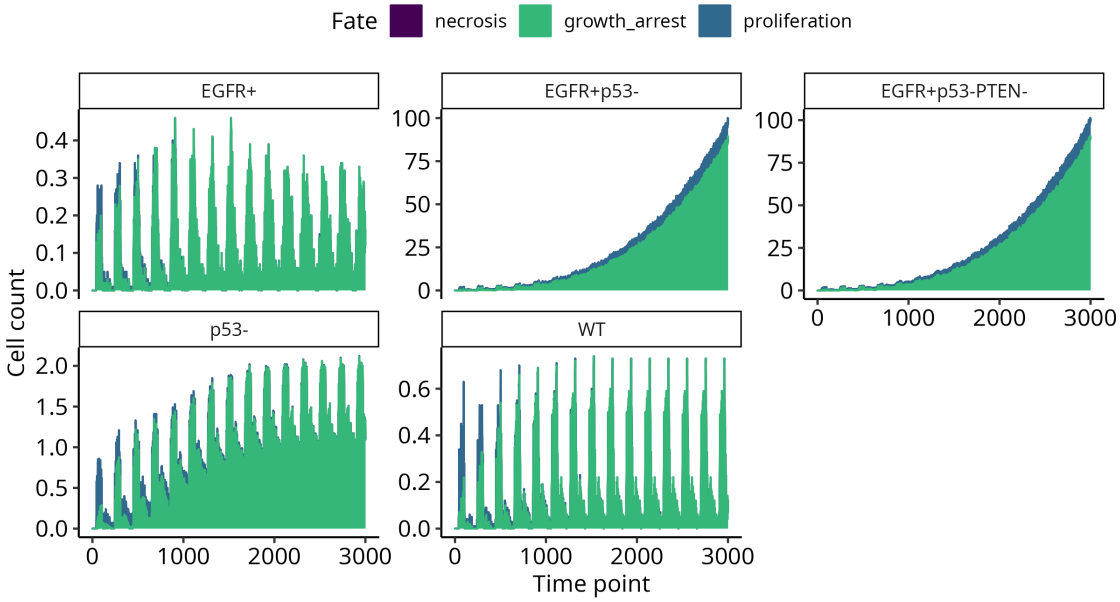
Figure 4.8 shows the fates activated by cells in the competition simulation, with apoptosis as a cumulative variable. There it can be seen that most cells activate apoptosis, and tumour growth is driven by a few cells that go into proliferation (see figure 4.8 b for a better view of the non-apoptosis fates). That is the expected outcome of apoptosis being executed immediately, while proliferating cells have to reach the end of the decision cycle without activating apoptosis or growth arrest. Still, although not pictured in figure 4.8, cells spend most of their time in a no-fate-reached state, due to apoptosis not being in the attractor for their network and initial state, and having to wait until the end of a decision cycle to activate growth arrest or proliferation (200 steps in these simulations).



Figure 4.7: Growth of a heterogeneous tumour at several time points. The data in this figure is from a different simulation to the one shown in figure 4.5, but uses the same parameters, with the exception that this one was run in 2-dimensional space with an hexagonal distribution, for the purposes of visualisation. This shows one replicate. Cells are coloured by their clonal identity. Dead cells (those that entered apoptosis) are not plotted since they are immediately removed from the simulation. Necrotic cells would appear in this figure, but with the given parameters there were none.



(a)



(b)

Figure 4.8: Fate activation during the competition simulation. At each time step, the number of cells activating a fate is represented here, coloured by fate. The cells that have not made a decision at that time point are not plotted. The data used here is from the same simulation as figure 4.5. In (a), the apoptosis fate is plotted as a cumulative variable, in order to represent the total number of cells that have died during the simulation. (b) omits the apoptosis fate for clarity. As in figure 4.5, this plot shows the average of 100 simulations, and because in most replicates 0 WT or EGFR+ cells expressed a fate other than apoptosis, the average is below 1.

4.2.3 The role of the microenvironment in the model evolution

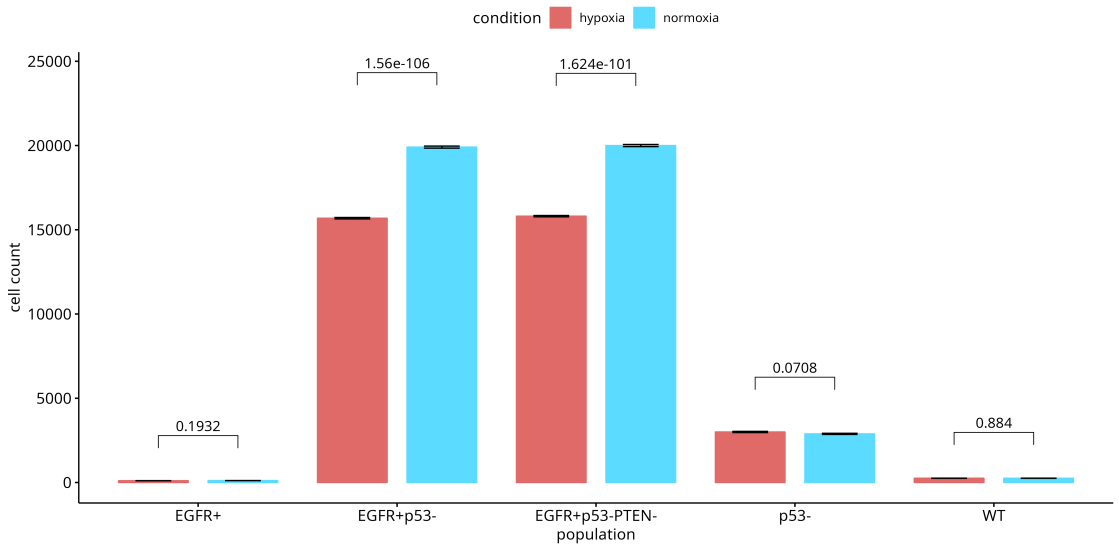


Figure 4.9: Growth comparison between normoxic and hypoxic monoclonal tumours. The figure shows the amount of of cells in each population at the end of the simulation (t=3000), both in hypoxic and normoxic simulations. The data shown are the means of 100 replicates for every simulation. A different simulation was run for every population and condition (normoxia or hypoxia). The error bars represent the standard error of the mean. A t-test was run for every pair of conditions per population — the numbers shown above the brackets are the Holm adjusted p-value for that pair. The normoxia data used for this figure and that used for figure 4.3 are the same.

An important part of the model, as described above, is the simulation of the tumour microenvironment. In *microCJ*, this comprises mainly the diffusion of substances such as oxygen, growth factors and drugs. I looked into the effect hypoxia has over the tumour model, making use of the hypoxia-related nodes added by Voukantsis *et al.*[257] to the MAPK network, as described in chapter 3.

Throughout the following experiments, oxygen levels are set at a high value of 0.28mM for normoxia and a low value of 0.04mM for hypoxia simulations and finally, The threshold for activation of the Oxygen_supply node is set to 0.022 mM, the physiological hypoxia[339].

Figures 4.9 and 4.10 show a comparison between simulations with the same parameters as figures 4.3 and 4.5, respectively, and other two simulations run with the same parameters but in a hypoxic environment (figure 4.11). One

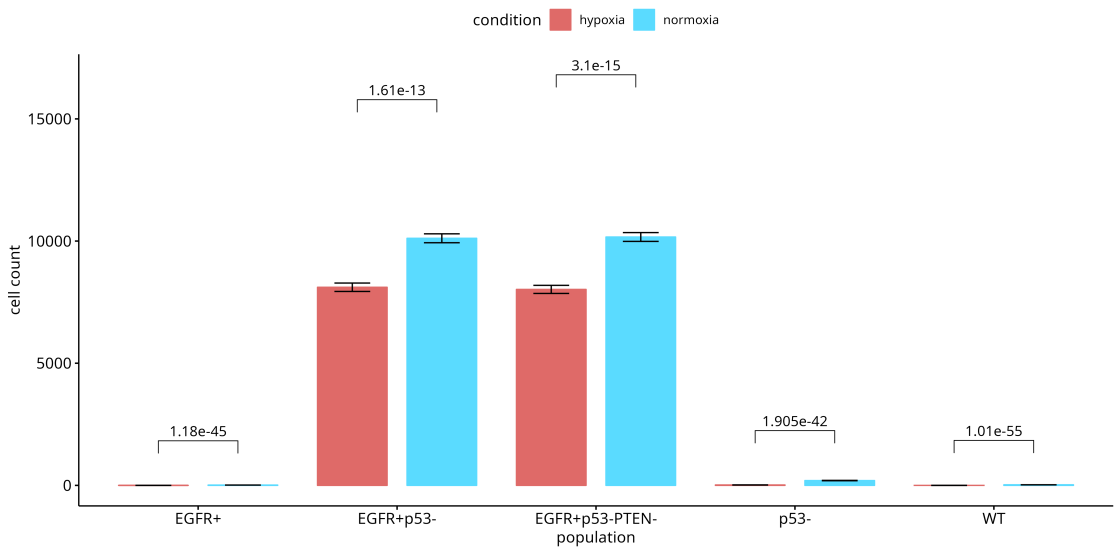


Figure 4.10: Growth comparison between normoxic and hypoxic multiclonal tumours. For legend explanation, see figure 4.9. The difference is that this figure shows the result of simulations with multiple subpopulations growing together in competition, as in figure 4.5, which plots the same normoxia data used here.

can see that hypoxia significantly hinders the growth of the populations in the simulation, especially in the case of the most aggressive ones (EGFR+p53- and EGFR+p53-PTEN-) whereas the difference is not substantial in the case of the slowest growing ones.

If one looks at the node expression profile for the hypoxic simulation (figure 4.12), there are no relevant predicted changes from the normoxic simulation. The reason for this is that with the network I use for these simulations (figure 3.3) there are only two outputs or readouts of oxygen concentration: necrosis and production of growth factors. Cells that go into necrosis are not represented in the figure, since their networks stop updating, and the production of EGF cannot be properly seen in a discrete time point heatmap. This is a limitation of the model presented here: it is known that hypoxia produces an array of modifications of the cell's behaviour, such as metabolic dysregulation, genomic instability, changes to motility or downregulation of DNA damage response genes. Most of these could be added to the model through additions to the GRN, and the changes in motility would be an interesting addition to a model of migration and metastasis, but these effects have not been considered in this model.

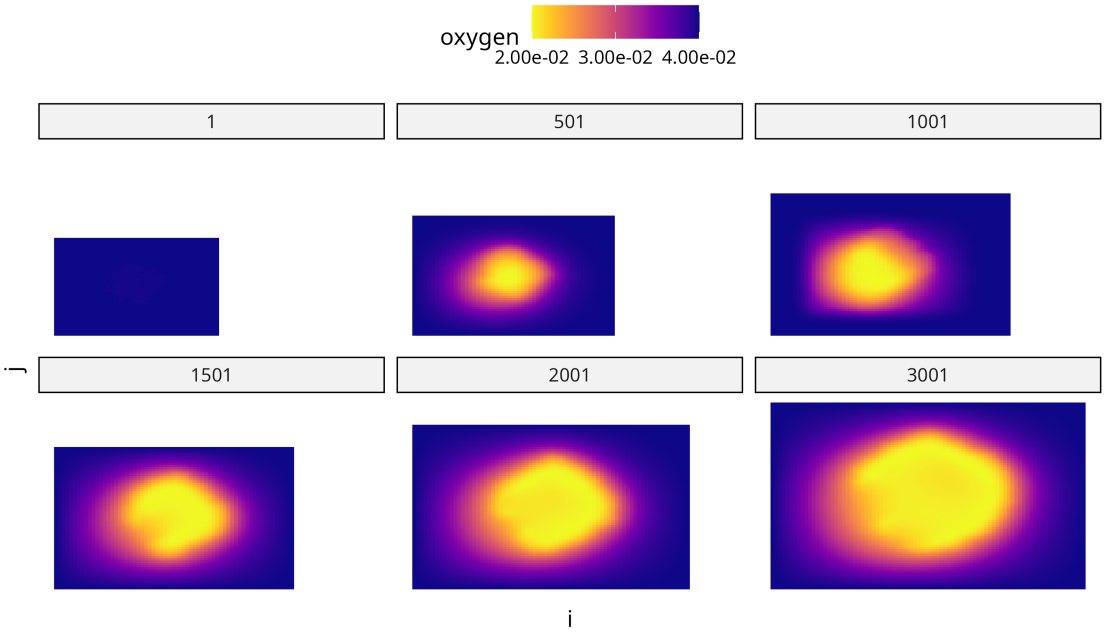


Figure 4.11: Oxygen diffusion through a 2D space hypoxic simulation. For several time points, the oxygen concentration in each voxel of the simulation space has been plotted as a heatmap, with tile colour representing concentration. This was a simulation run with competing populations in cells, plotted in figures 4.13 and 4.14, and their oxygen consumption can be seen in the oxygen depleted area in the center of the space (bright yellow tiles). The space grows in size as the tumour gets bigger, maintaining a constant distance between the boundary and the outermost cells, to expedite computation.

Figures 4.13 and 4.14 show a 2D view of an hypoxic simulation with competing subpopulations. The formation of a necrotic core can be seen, and how the only cells that manage to escape it are from the two most aggressive subpopulations (EGFR+p53- and EGFR+p53-PTEN-).

After looking at hypoxia, I turn to EGF diffusion. In figure 4.15 one can see the predicted proliferation inducing effects of allowing cells to produce EGF when the corresponding node is activated in the network. There is not much of a difference for WT and PTEN- cells, since the binding of EGF by EGFR is not enough for them to consistently proliferate if p53 and its downstream are active, as is also the case for non EGFR+ mutants. But for p53- and p53-PTEN- subpopulations the difference is quite significative. Figure 4.16 highlights that the production of EGF is carried out by just a few cells that present the relevant network state for it, and

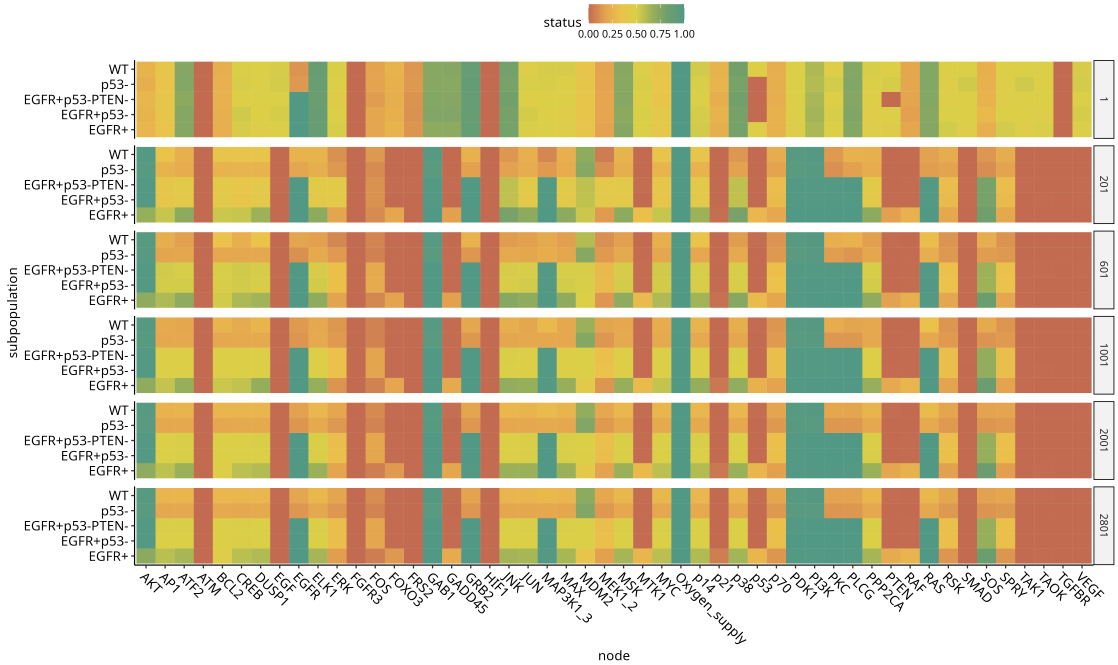


Figure 4.12: Average node activation of independent populations across different time points under hypoxia. For figure explanation see figure 4.4. Same data used as the hypoxia data in figure 4.9

the factor is quickly consumed by the producer cell itself and its neighbours.

Production of EGF is predicted to offer a competitive advantage during hypoxia, as seen in figure 4.17, where the cells that proliferate fast enough to escape the necrotic core are able to produce EGF and thus grow at a faster rate than without EGF in the microenvironment.

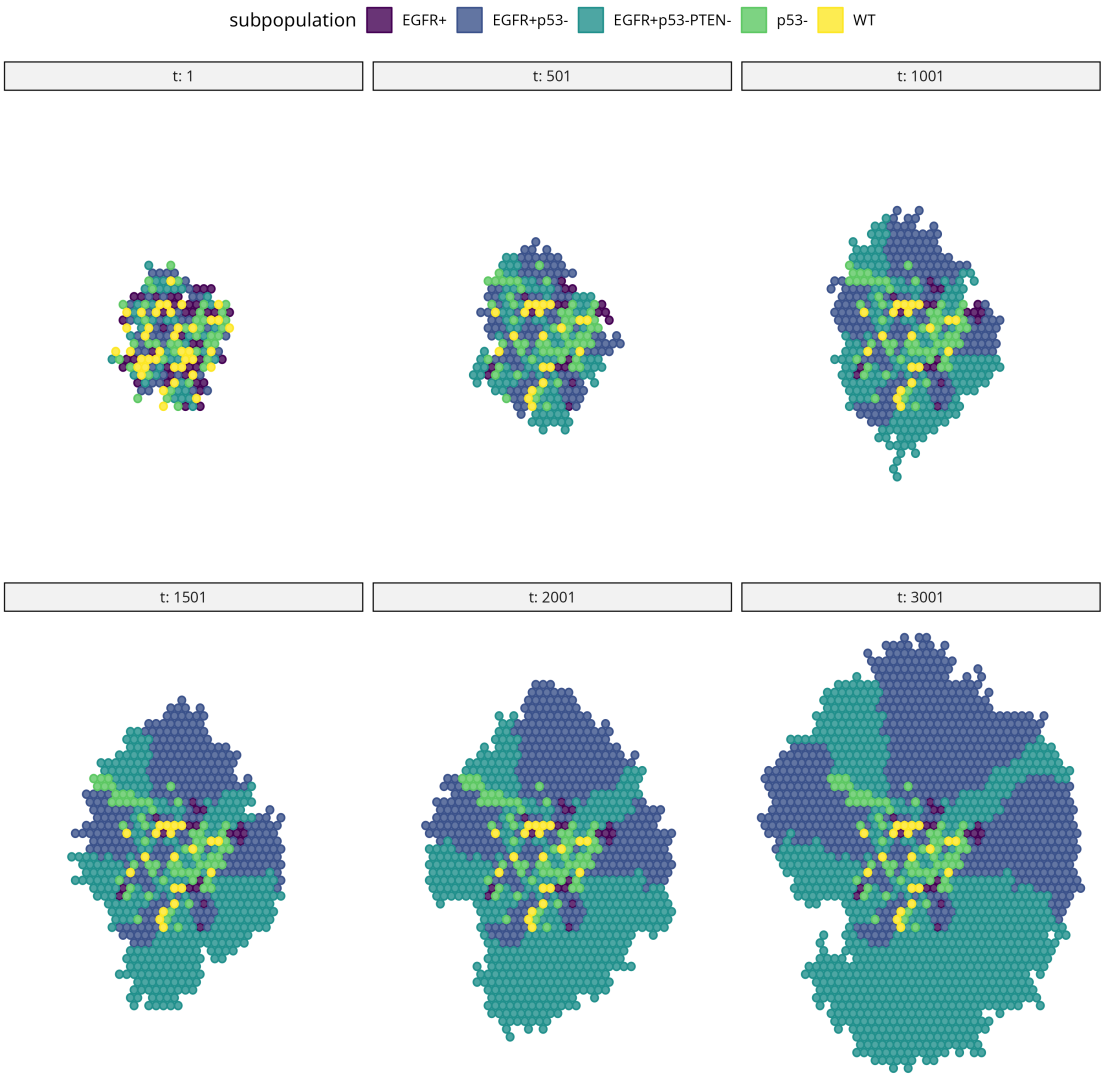


Figure 4.13: Growth of a heterogeneous tumour under hypoxic conditions at several time points. For an explanation of the figure, see figure 4.7. The data for this figure are the same as used for figure 4.11

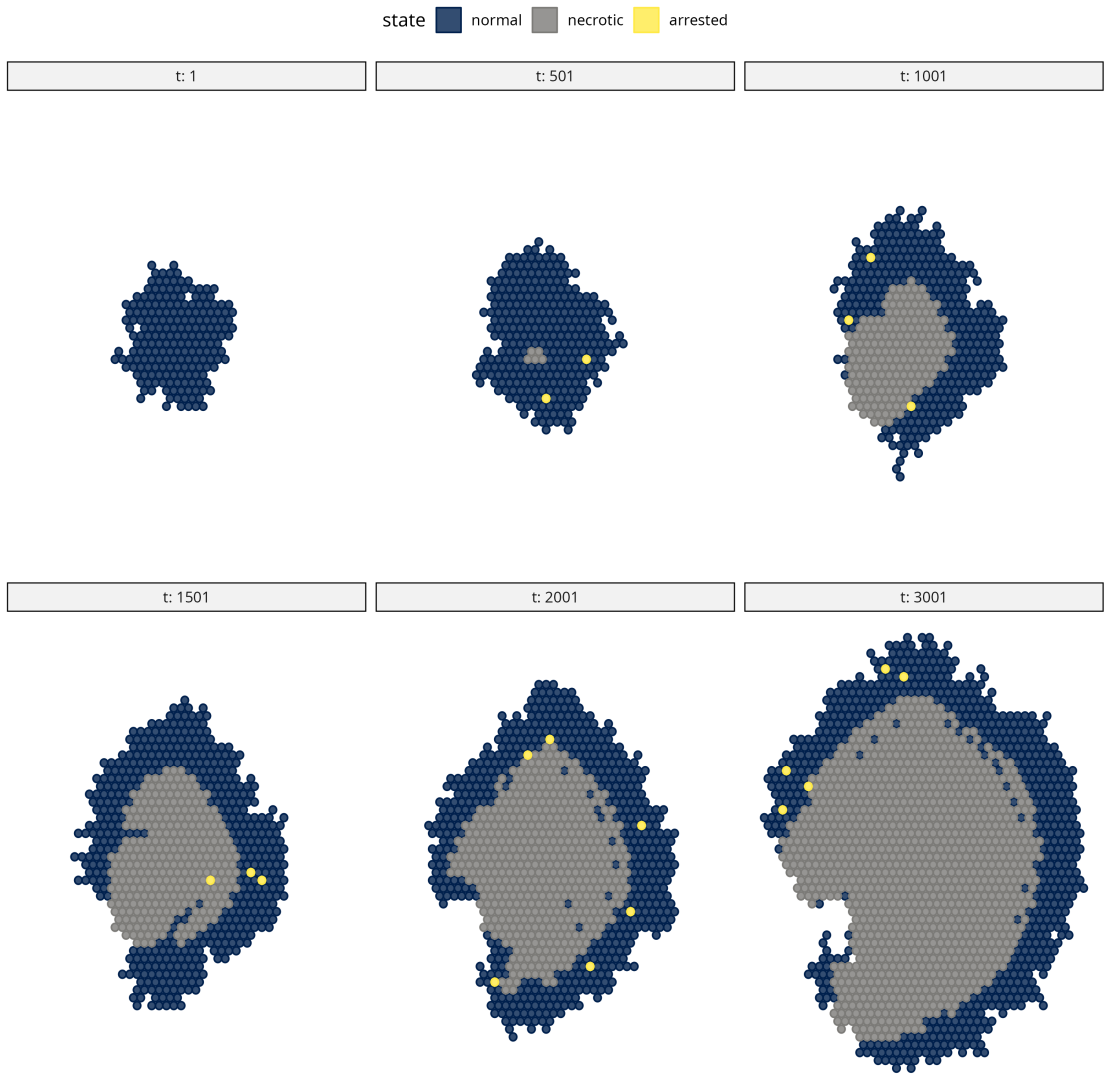


Figure 4.14: State of each cell in a heterogeneous tumour under hypoxic conditions at several time points. Same simulation as in figure 4.13, but here colour represents the state of the cells, one of normal, necrotic or arrested (cells that are in growth arrest). Apoptotic cells are not plotted as they are removed from the simulation as soon they activate the apoptosis fate. The formation of the necrotic core can be seen (grey mass at the center of the figure).

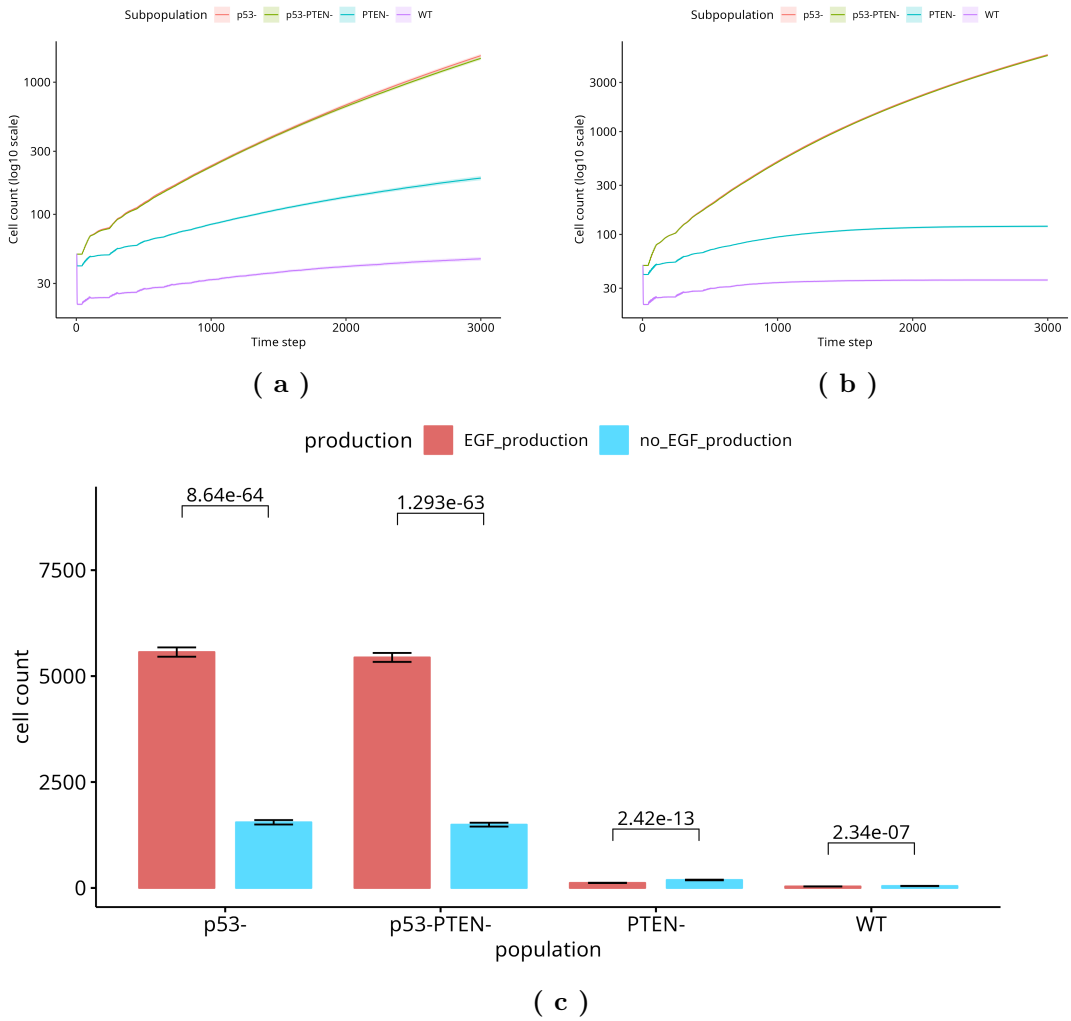


Figure 4.15: EGF diffusion. The plots show the results of simulations with (a): no production of EGF, versus (b): production of EGF. The subpopulations were run in competition, and were p53-, PTEN-, P53-PTEN-, and WT. (c) shows a comparison between both options. The data shown are the means of 100 replicates for every simulation. The error bars represent the standard error of the mean. A t-test was run for every pair of conditions per population — the numbers shown above the brackets are the Holm adjusted p-value for that pair.

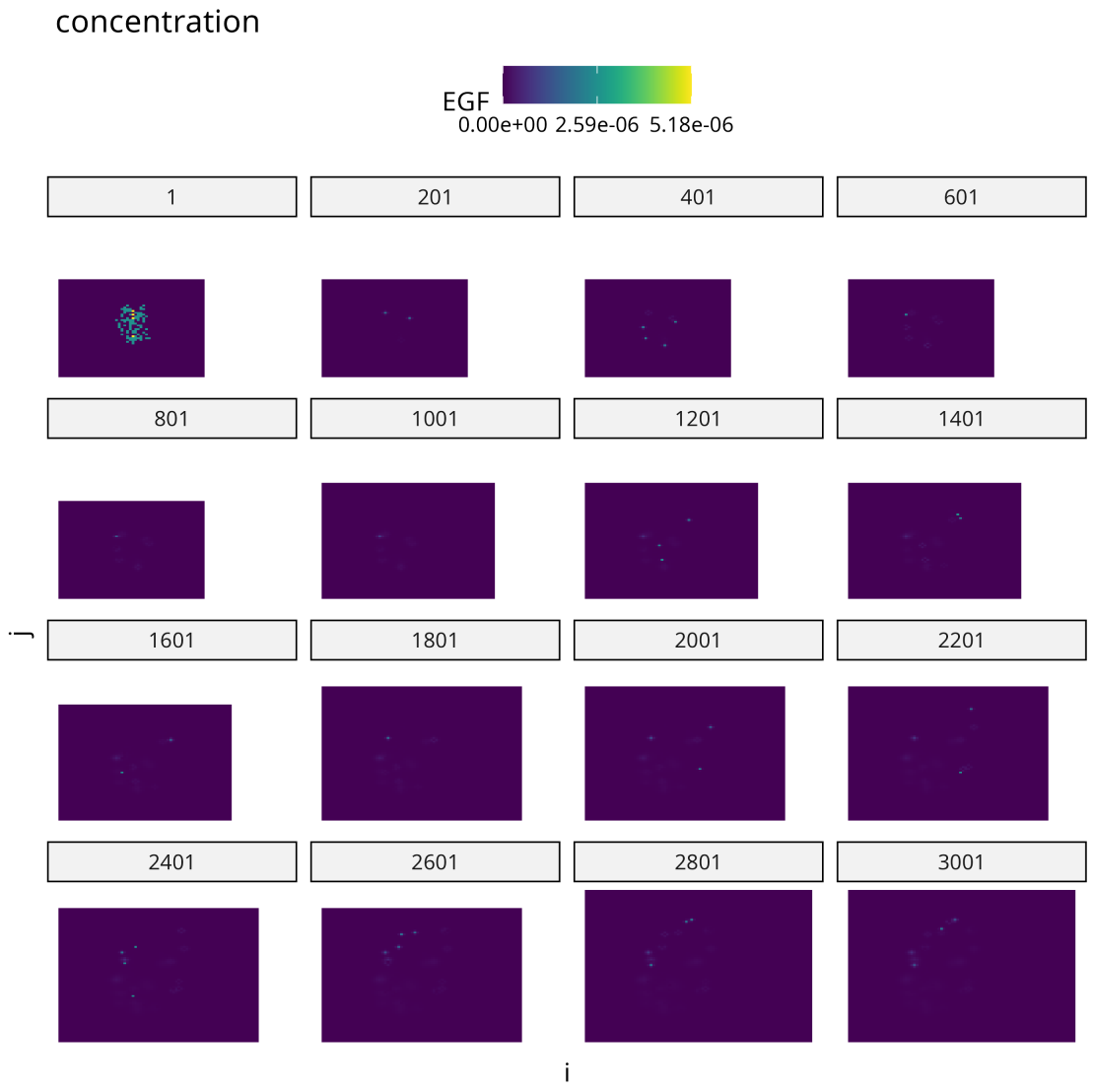
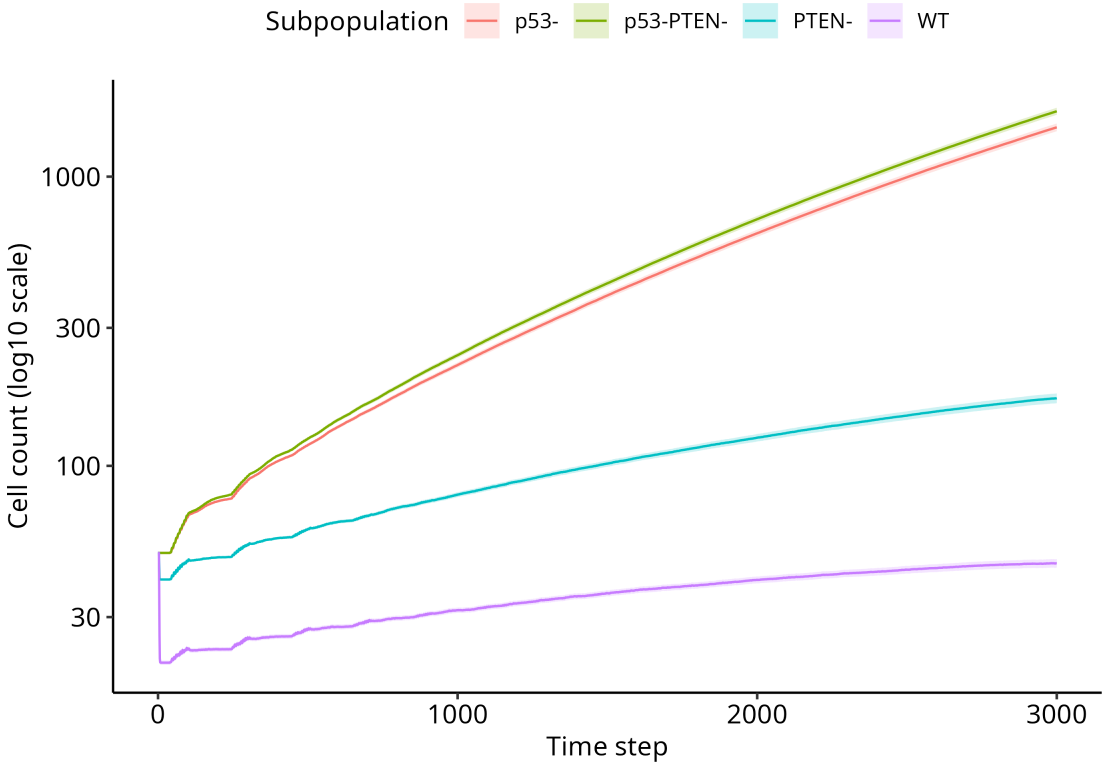
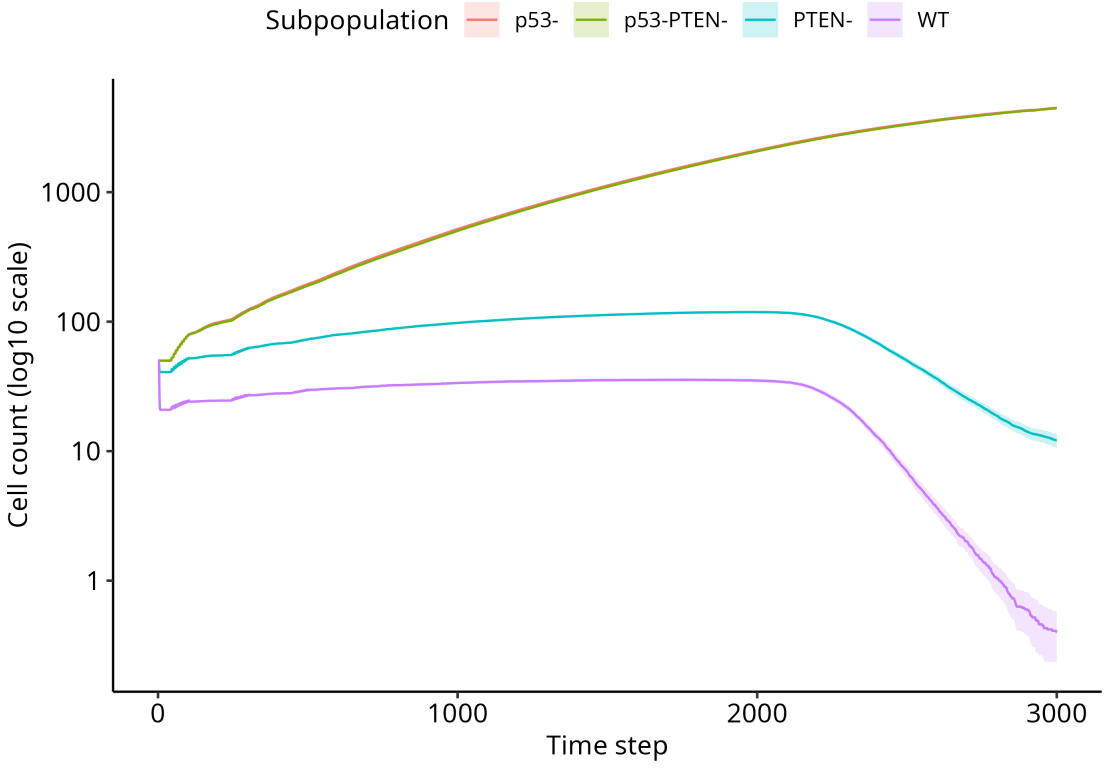


Figure 4.16: EGF diffusion through a 2D space simulation. For an explanation of the figure, see figure 4.11



(a)

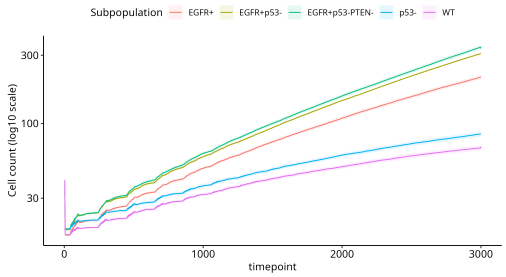


(b)

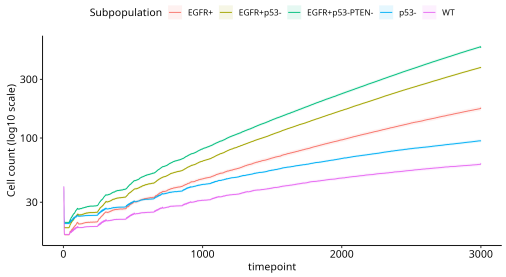
Figure 4.17: Effect of EGF production in hypoxic conditions. The plots show the results of simulations with (a): no production of EGF, versus (b): production of EGF; both with hypoxic conditions (figure 4.11). The subpopulations were run in competition.

4.2.4 Probabilistic Boolean Networks for a more granular simulation

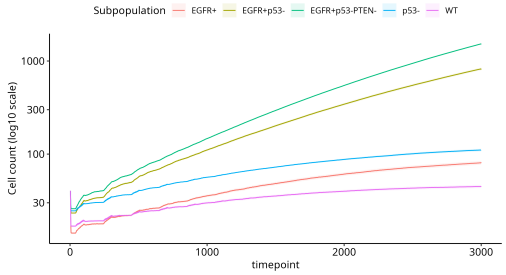
To test the probabilistic mutations feature of *microCJ*, I ran the same competition simulation as in figure 4.5, but using PBNs instead of a synchronous updating strategy (figure 4.18). Thus, mutations are not simply a fixed one or zero for the whole duration of the simulation, but probabilities for nodes to be active or inactive (see Features - Network Modelling section above). This way, more granularity is possible, allowing *microCJ* to model different levels of over-expression, for example. This forms the basis for *Ingres* integration into *microCJ*, as it will be explored in chapter 5. figure 4.18 demonstrates that even at the lowest probabilities — e.g. a 5% over or under-expression — some differences can be seen between subpopulations. As the mutation value increases, these predicted differences become more marked, as does the total number of cells after the 3000 steps — from a few hundreds to over 12,000 cells in the most aggressive population, EGFR+p53-PTEN-. This proves that, albeit a simple mechanism, the inclusion of probabilistic mutations can clearly model differences between populations with slightly different levels of over- or under-expresssion, providing support for the soundness of using such a mechanism to fit *microCJ* with transcriptomics data.



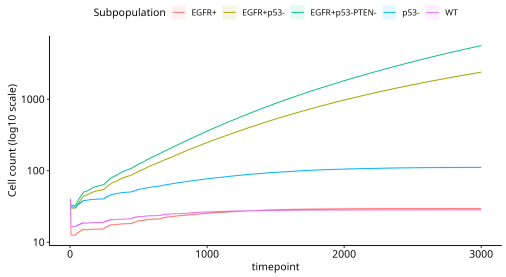
(**a**) p = 0.05



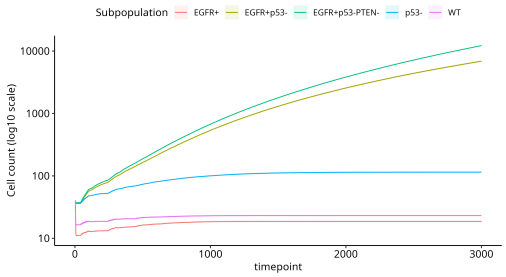
(**b**) p = 0.10



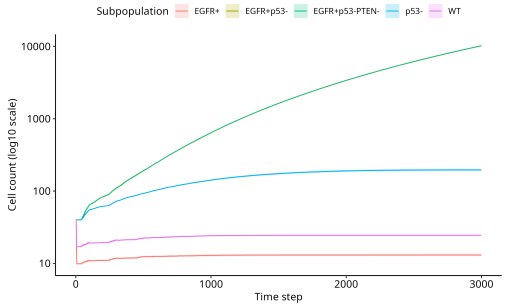
(**c**) p = 0.25



(**d**) p = 0.50



(**e**) p = 0.75



(**f**) p = 1

Figure 4.18: Competition simulations using probabilistic networks. These simulations used the same parameters as in figure 4.5, but instead of synchronous networks, PBNs were used. The PBN used is identical to the synchronous one figure 3.3 but mutations are modelled as additional rules for the mutated node, as explained before. In the case of the EGFR+p53- subpopulation, in (a), the node EGFR has a probability p=0.05 of being fixed to 1, and a p=0.95 of functioning normally of 0.95. Conversely, p53 has p=0.05 of being fixed to 0. The figure shows five different simulations with increasing probabilities for the mutations. figure 4.18 f shows figure 4.5 for comparison.

4.2.5 Acquired mutations and clonal evolution

An alternative strategy to use *microCJ* and gain mechanistic insights of the object of study is to let the model to evolve on its own, rather than predefining the subpopulations and their mutations. Through the use of the *acquired-mutation-probability*, it is possible for the cells in the model to gain new mutations. Since the cells are subject to selective pressure — non-beneficial mutations would cause the cell to go into apoptosis and therefore to be unable to pass down that mutation — by studying which genes have which kind of mutation at the end of the simulation one can see which ones are the ones that give cells the bigger evolutionary advantage. I ran two such analyses, one as a control and another one with cells receiving sustained DNA damage as a selective pressure. Both simulations had 100 replicates each and were run for 10,000 time steps to give the cells enough time to evolve and be selected. The number of initial cells was set to 500. In both simulations, the probability of a cell acquiring a new mutation at each time step was set to 0.8. The results of these experiments are illustrated in figure 4.19, where all the mutation values from all replicates (around 10^6 cells) at time step = 10,000 are shown as box plots, grouped by genes. These values are the probability for each gene to be fixed to one at each time step, in the case of a positive mutation. A negative value means that that mutation is for under-expression.

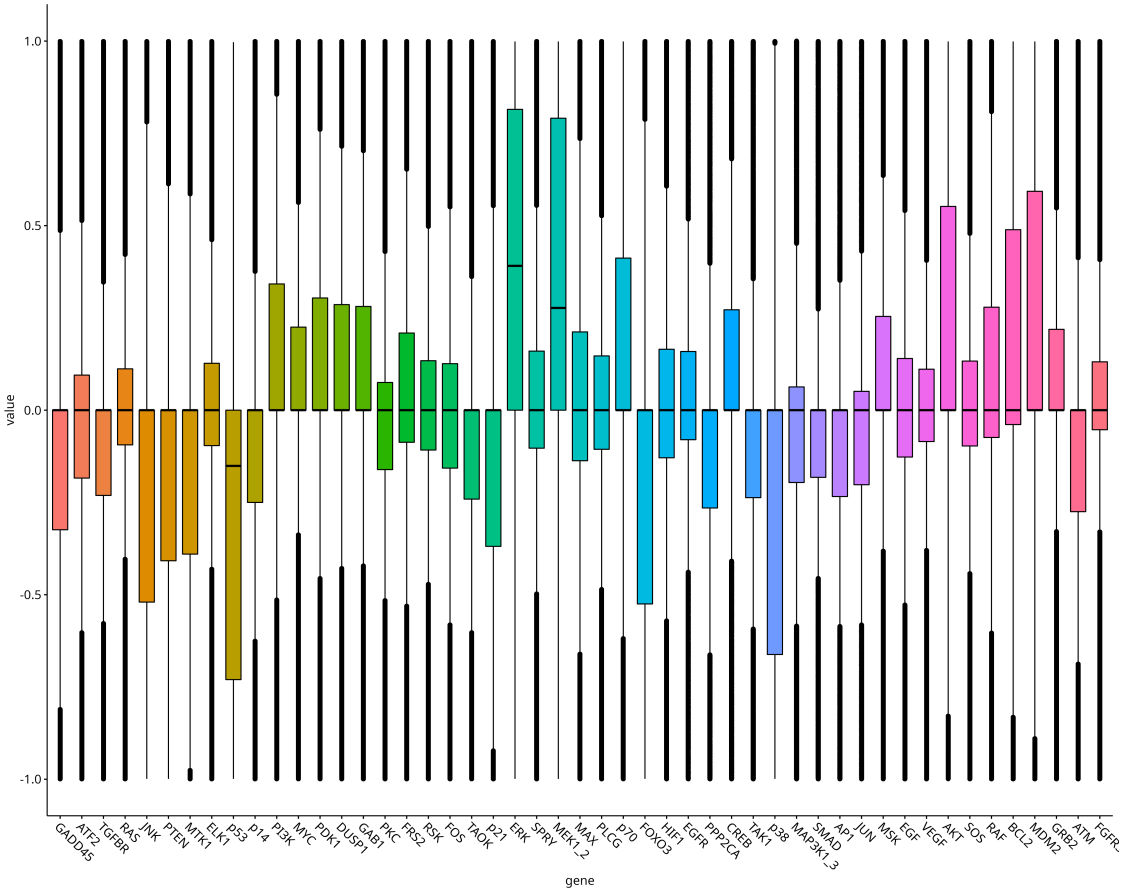
I assumed that mutations present in the surviving cells at the end of the simulation would have been selected by their superior fitness or proliferative advantage. Indeed, one can see that some genes are consistently mutated in a specific direction, such as p53-, PTEN-, ERK+, MYC+ or MEK+. These mutations are consistent with what one observes in real life tumours. I defined the consistently mutated genes as those that had either of the first or third quartile = 0, as that would mean that most of the mutations share their sign, that is, the majority of cells among independent replicates present over or under expression. I then plotted the data for those genes in a comparison between the control and the DNA damage simulations, shown in figure 4.20. Genes that are frequently mutated in resistant tumours are

significantly more over- or under- expressed here as an effect of DNA damage, such as p53-, PTEN-, ERK+, MEK+ or AKT+ (all with ($p<10^{-4}$).

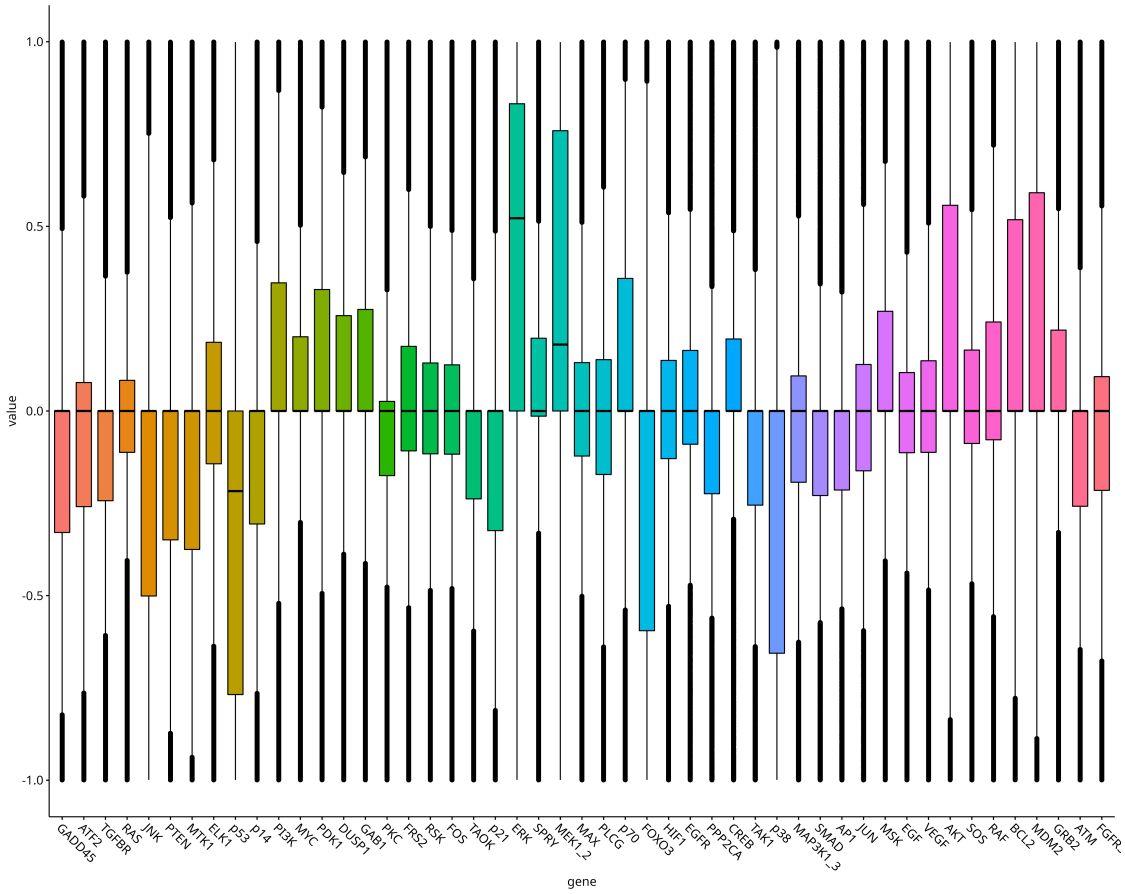
A more precise definition of advantageous mutations would be those present in cells with a higher division rate. A readout for this in *microCJ* would be how many times a given cell has divided at the end of the simulation, thereon named the cell *generation*. The cells with the highest generation at the end of the simulation are deemed to be the most succesful ones, and so in figure 4.21 I look at the correlation between generation and gene mutation. p53 and p21 are the two genes more negatively correlated with generation: their under-expression is advantegeous. ERK is the most positively correlated gene. Interestingly, proliferation-involved genes such as MYC, AKT or ERK are negatively correlated with tumour suppressing genes such as p53 or p21, showing that they are often mutated in opposite directions together in these highly proliferative cells.

Finally, figures 4.22 and 4.23 show the cells at the end of the simulation coloured by each gene mutation and by generation, respectively. One can observe that areas of the tumor with a higher overall generation, meaning that cells have a higher proliferative rate there, match the areas where ERK has an over-expressing mutation (bottom-left area), or where there is MEK+, PTEN- and TGFBR- (top-right area) for example. As it can be seen from the shape of these areas, most of them originate with a single cell that acquires that mutation and then rapidly takes over the space around them in successive divisions. One can also see that non-advantageous mutations sometimes appear and are carried on for some time, but generally die off shortly after, as in the little AKT- area in the middle right part of the tumour. On the other hand, there are also passenger mutations, non-advantageous alterations that form part of a successful subpopulation of the tumour: they are not beneficial by themselves, but they coexist with advantageous mutations that give the cell in question an overall boost to proliferative rate (driver mutations). This is the case, for example, in the top right area, which despite being very proliferative has a PI3K- mutation.

This shows that *in silico* WT cells, when subjected to genomic instability (a high mutation rate) evolve and develop mutations commonly seen in TNBC and other cancer types, which could be further tested with clinical data and used to predict which mutations will arise as a result of a given set of selective pressure, such as therapy.


1.0
0.5
0.0
-0.5
-1.0
value
GADD45
ATF2
TGFBR
RAS
JNK
PTEN
MTK1
ELK1
P53
P14
PI3K
MYC
PDK1
DUSP1
GAB1
PKC
FRS2
RSK
FOS
TAOK
P21
ERK
SPRY
MEK1_2
MAX
PLCG
P70
FOXO3
HIF1
EGFR
PPP2CA
CREB
TAK1
P38
MAP3K1_3
SMAD
AP1
JUN
MSK
EGF
VEGF
AKT
SOS
RAF
BCL2
MDM2
GRB2
ATM
FGFR
gene

(a)



(b)

Figure 4.19: The acquired mutations during a simulation with sustained DNA damage, with a control simulation. Two simulations were run with 500 initial cells, all of them WT, and the acquired-mutation-probability parameter set to 0.8. (a) shows the control simulation, and (b) the sustained DNA damage one. The figure summarises the data points from all cells in the 100 replicates. In the box plots, the mid line is the median value, the lower and upper hinges correspond to the first and third quartiles, and the whiskers extend to the largest and smallest value, respectively, no further than 1.5 * inter-quartile range. Data beyond the end of the whiskers are plotted individually.

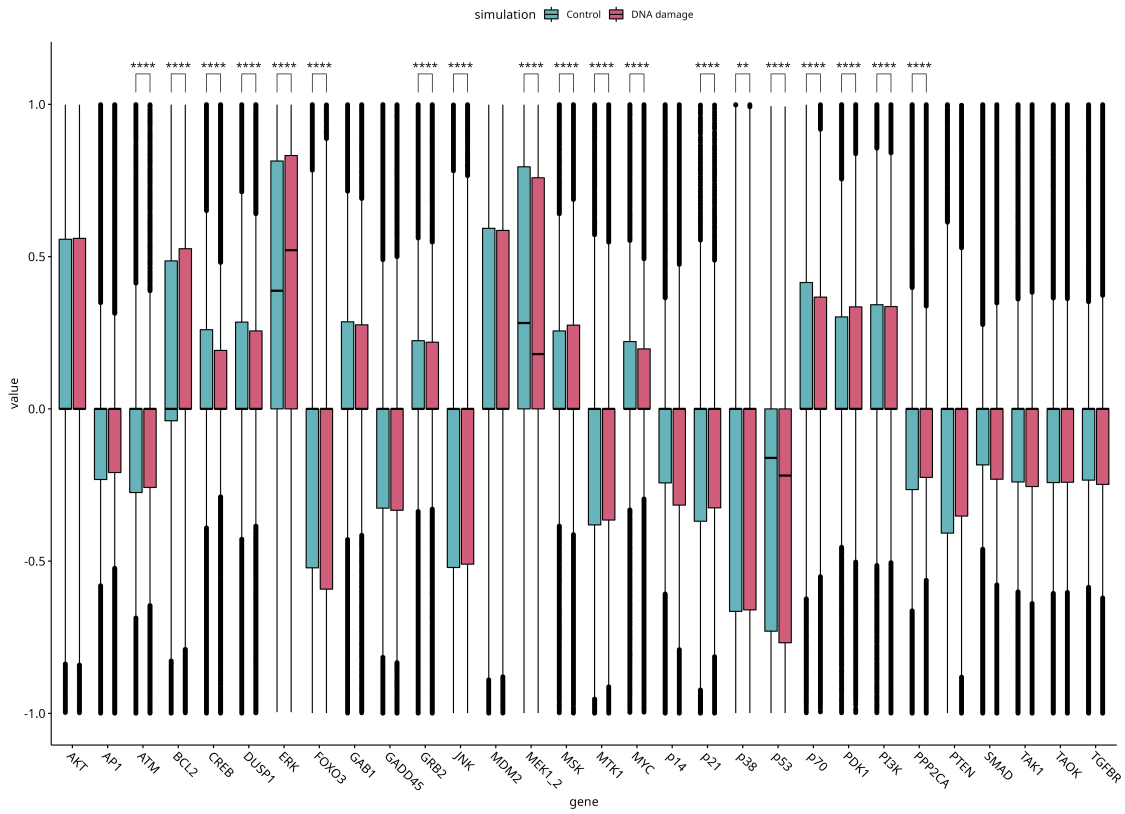


Figure 4.20: Comparison between mutated genes in control and sustained DNA damage simulation. Data from the same simulations as figure 4.19. The genes with either the first or the third quartile=0 were selected, they are the ones presenting consistent over- or under-expression. The figure shows the data points from both the control and the DNA damage simulations side by side, grouped by gene. Each pair was compared using a Wilcoxon test, and the p-value was adjusted using Holm's method. The legend for the symbols above the brackets is ****: $p<10^{-4}$; ***: $p<0.001$; **: $p<0.01$; *: $p<0.05$

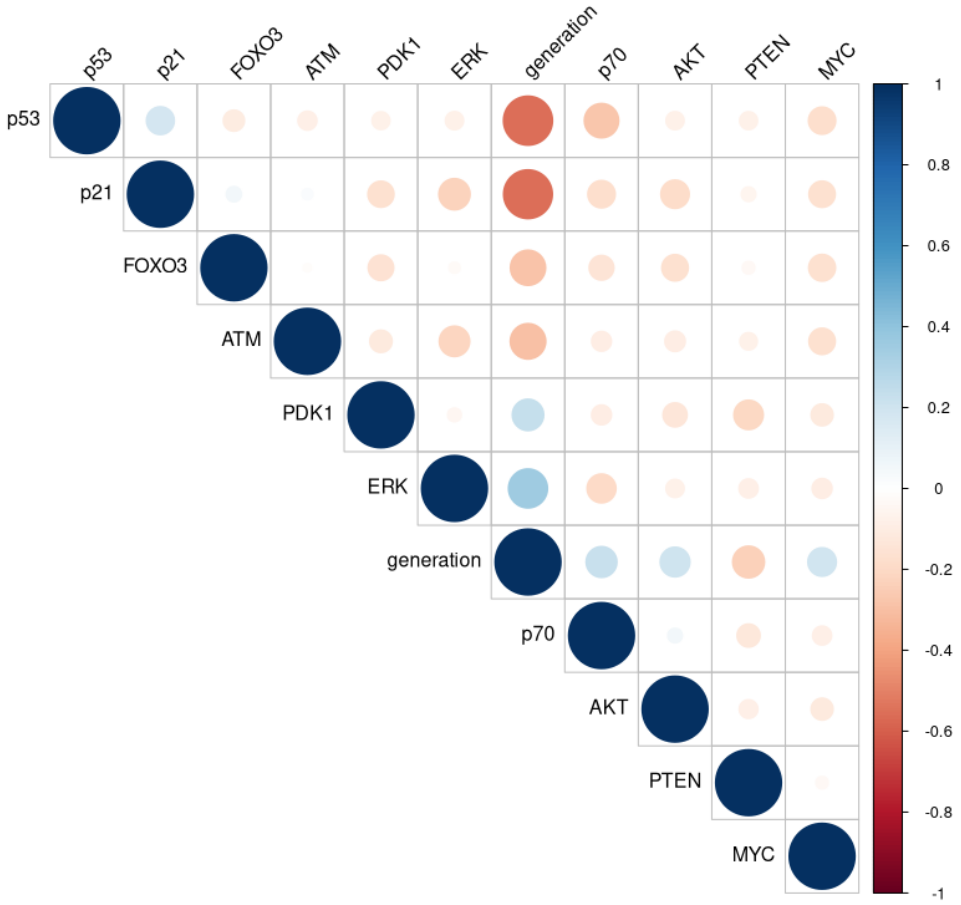
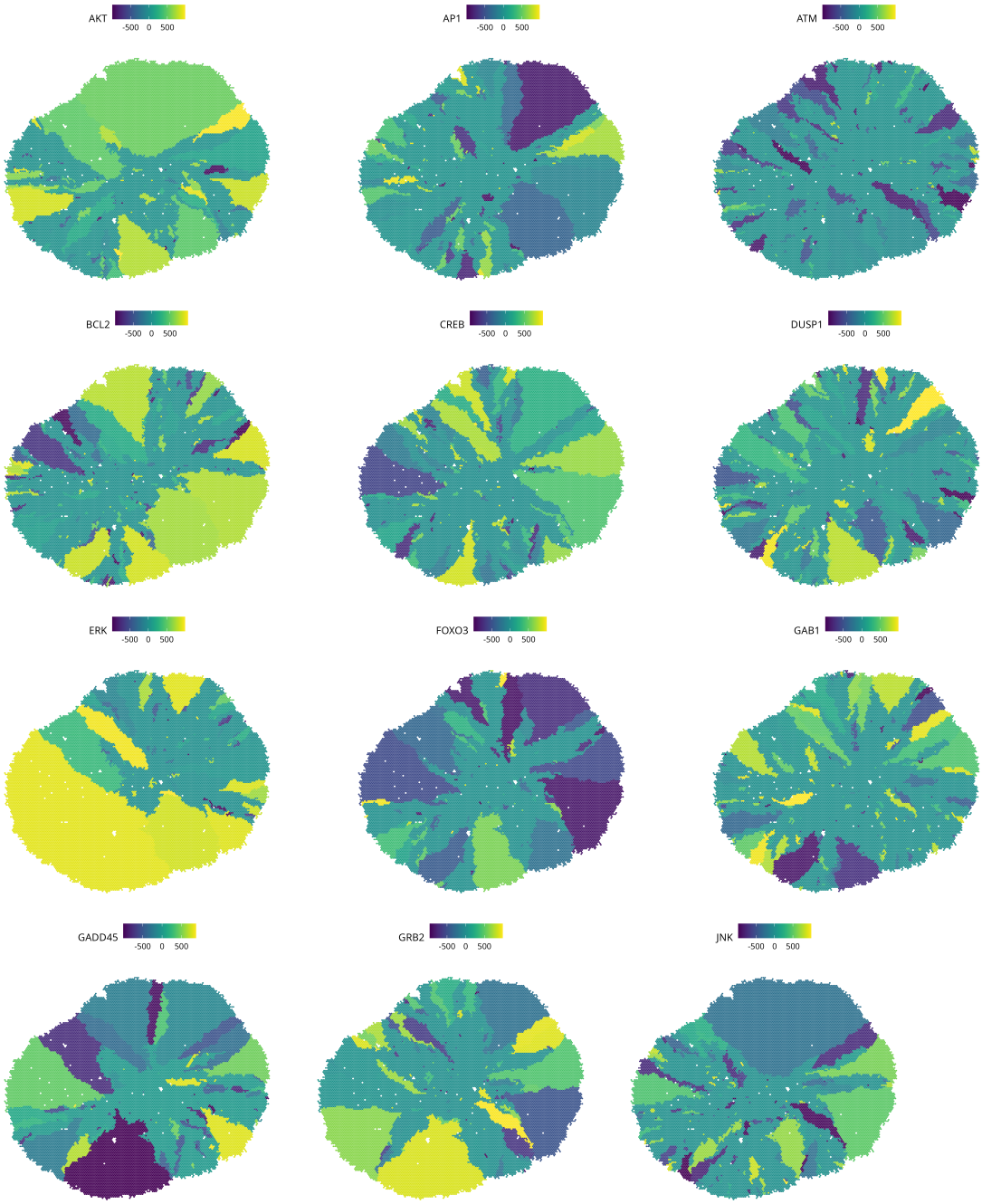
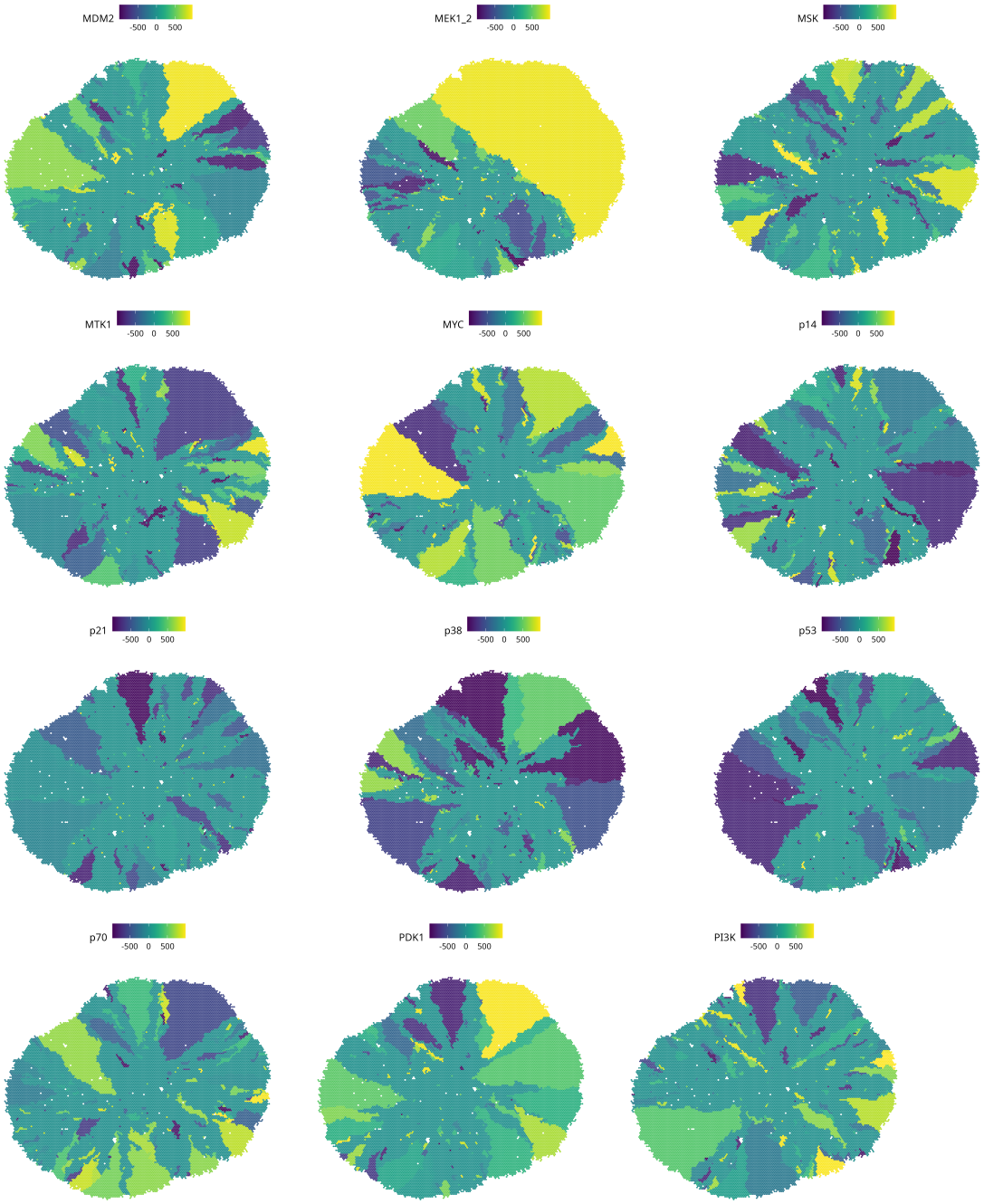


Figure 4.21: Correlation between mutated genes in the DNA damage simulation and cell generation. Cell *generation* is defined as the number of divisions they are a result of. The initial cells in the simulation will have generation=0, their daughters generation=1, and so on. Only the five most positively correlated and the five most negatively correlated genes are shown here. The size, transparency and colour of the circles is relative to the correlation.


AKT
-500 0 500
AP1
-500 0 500
ATM
-500 0 500
BCL2
-500 0 500
CREB
-500 0 500
DUSP1
-500 0 500
ERK
-500 0 500
FOXO3
-500 0 500
GAB1
-500 0 500
GADD45
-500 0 500
GRB2
-500 0 500
JNK
-500 0 500


MDM2
-500 0 500
MEK1_2
-500 0 500
MSK
-500 0 500
MTK1
-500 0 500
MYC
-500 0 500
p14
-500 0 500
p21
-500 0 500
p38
-500 0 500
p53
-500 0 500
p70
-500 0 500
PDK1
-500 0 500
PI3K
-500 0 500

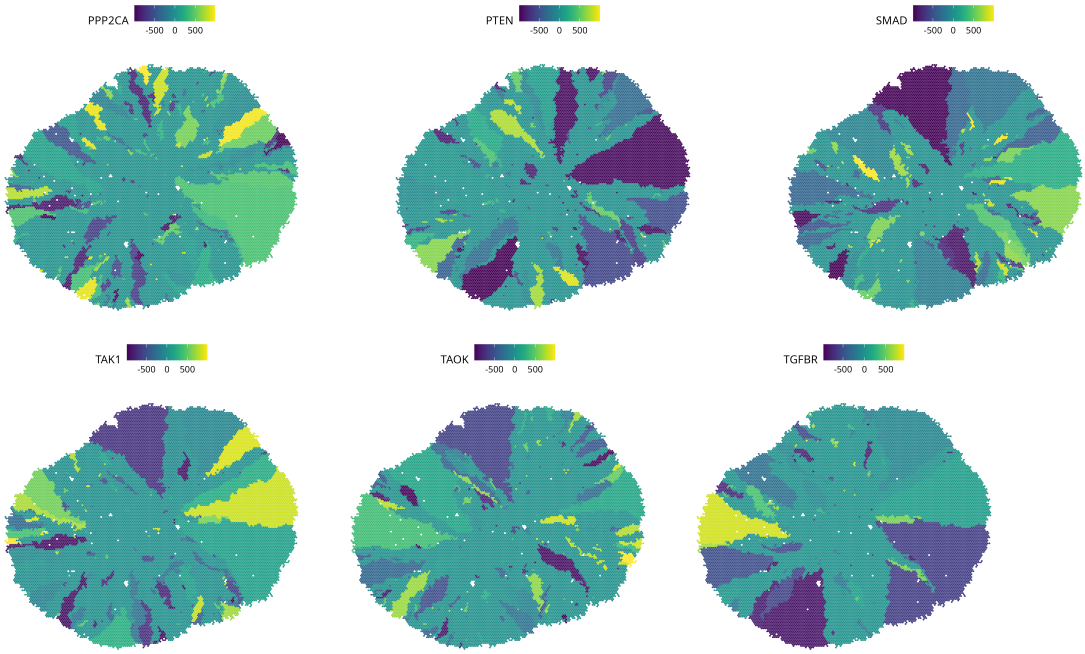


Figure 4.22: Mutations acquired by cell during a simulation with ongoing DNA damage. Continued from previous pages. Each of the plots correspond to one of the genes from figure 4.20. Here only one of the replicates, chosen at random, is plotted, with the data from t=10,000. Cells are coloured by the intensity of the mutation in question (its probability) and its direction: whether it is over-expression or under-expression.

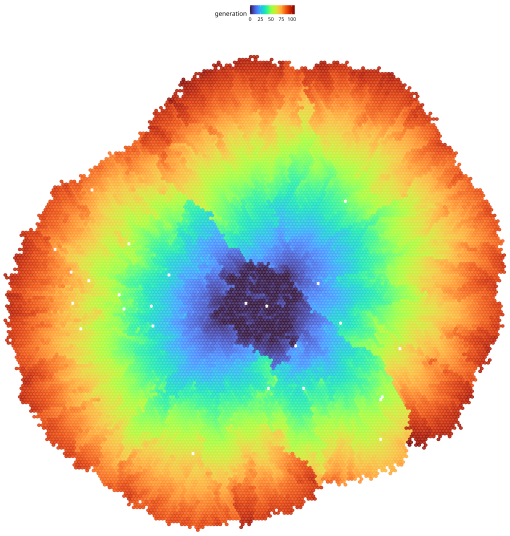
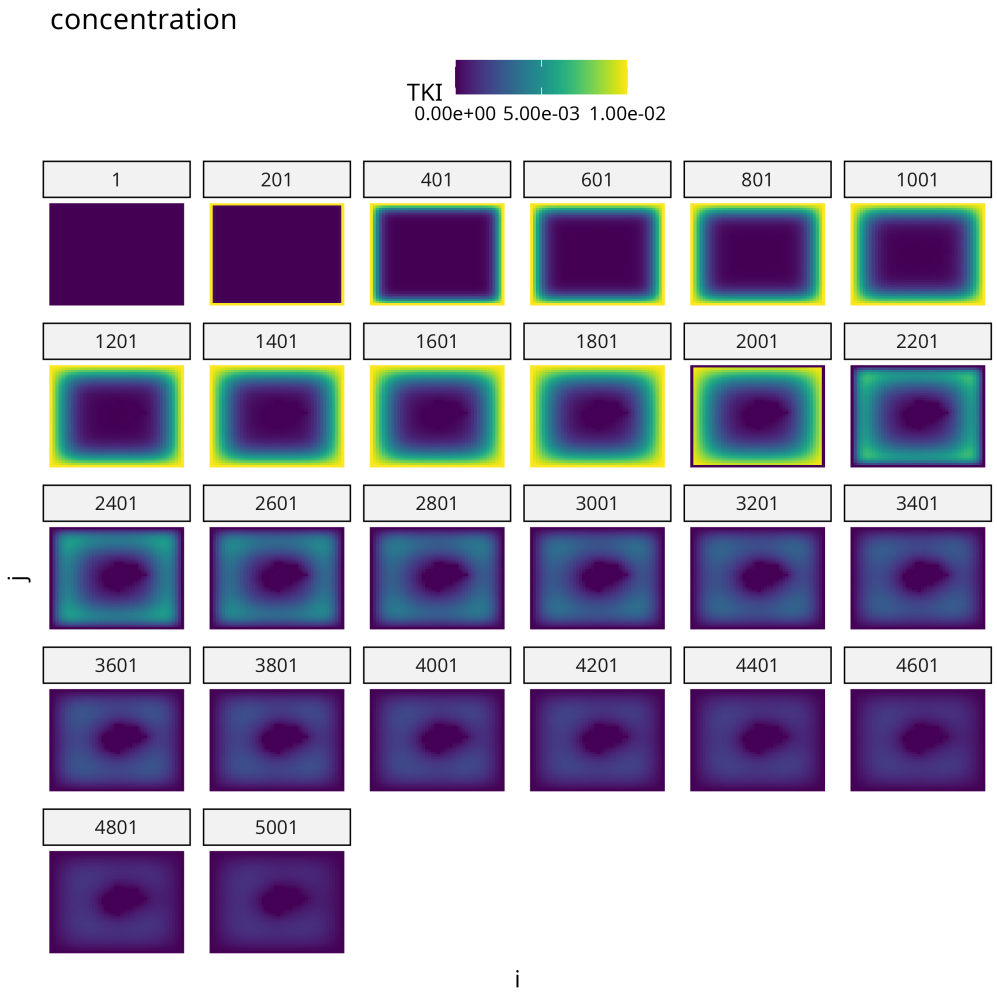


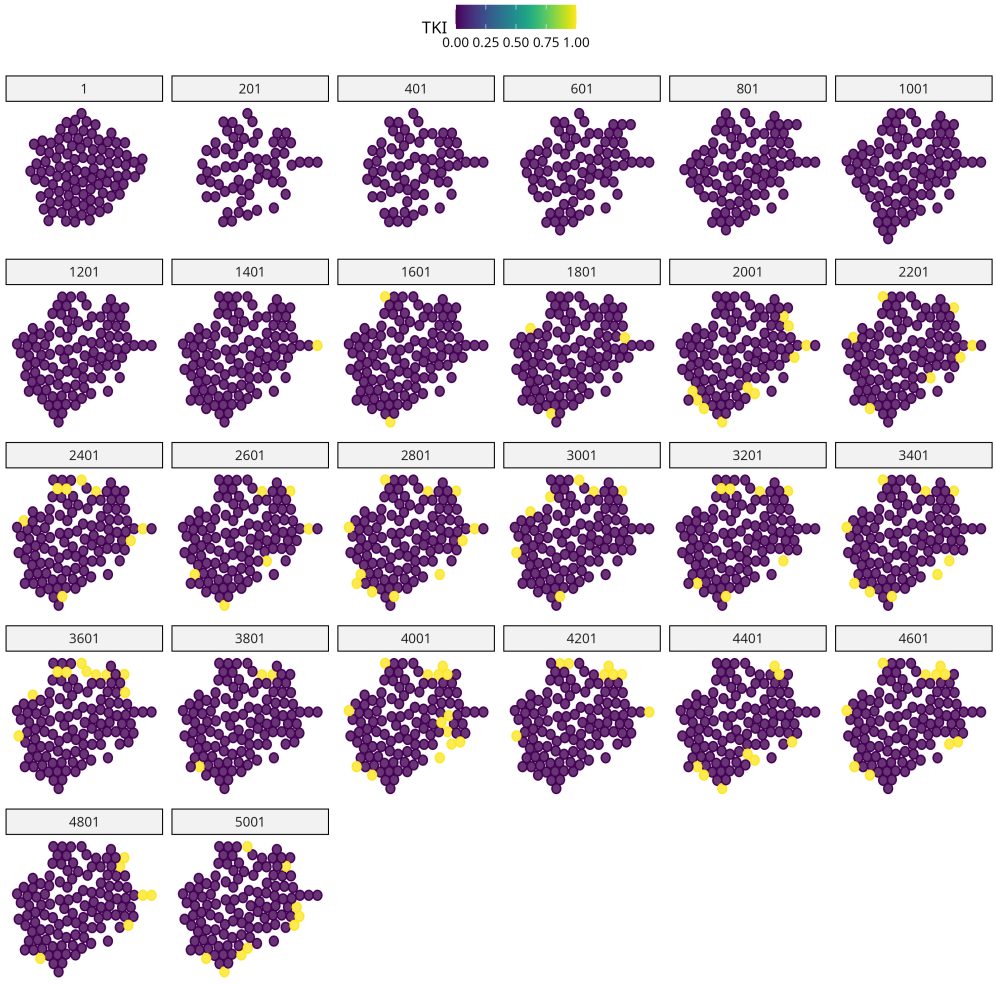
Figure 4.23: Cells in the DNA damage simulation, coloured by generation. Cell plot with data from the same simulation as figure 4.19 b . The cells are coloured by their *generation*. This visualisation allows to observe in which areas of the tumour the cells have divided more often.

4.2.6 Drug treatment

The last feature described in this chapter is drug treatments. I implemented them as substances that diffuse from the boundary of the simulation (like oxygen does, for example) but only during a specified period of time. Before and after the specified start and end point there is no drug coming from the boundary. When a cell consumes the drug — that is, the drug reaches the same voxel the cell is in and its concentration goes above a given threshold, a specific node becomes activated in that cell's network, as with any other diffusible substance in *microCJ*. That node is assumed to be inhibiting some other node in the network, simulating the action of the drug. To test this feature, I added a node to the network called "TKI" which inhibits the EGFR node, thus modelling the action of an EGFR TKI such

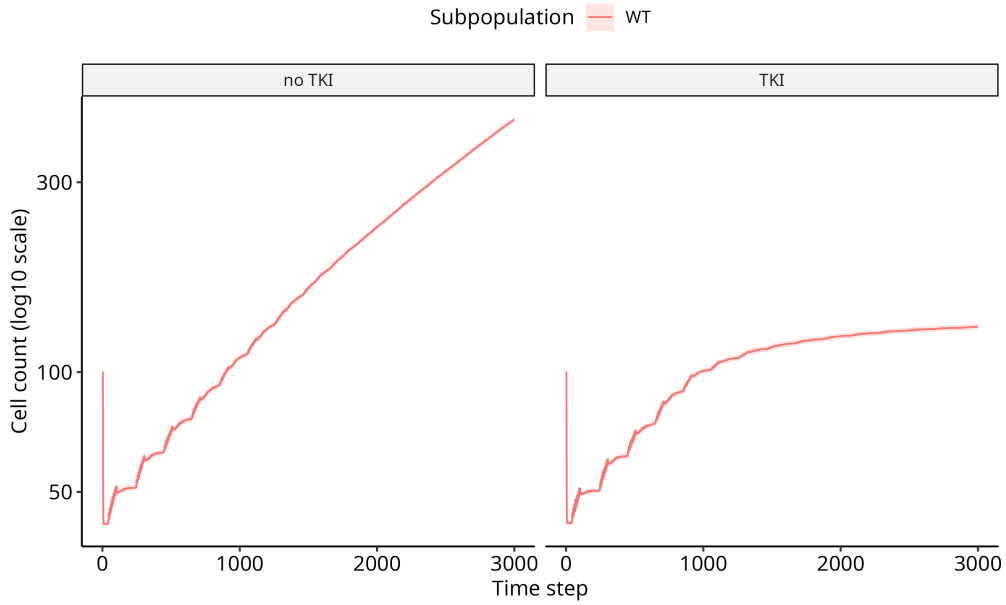


(a)

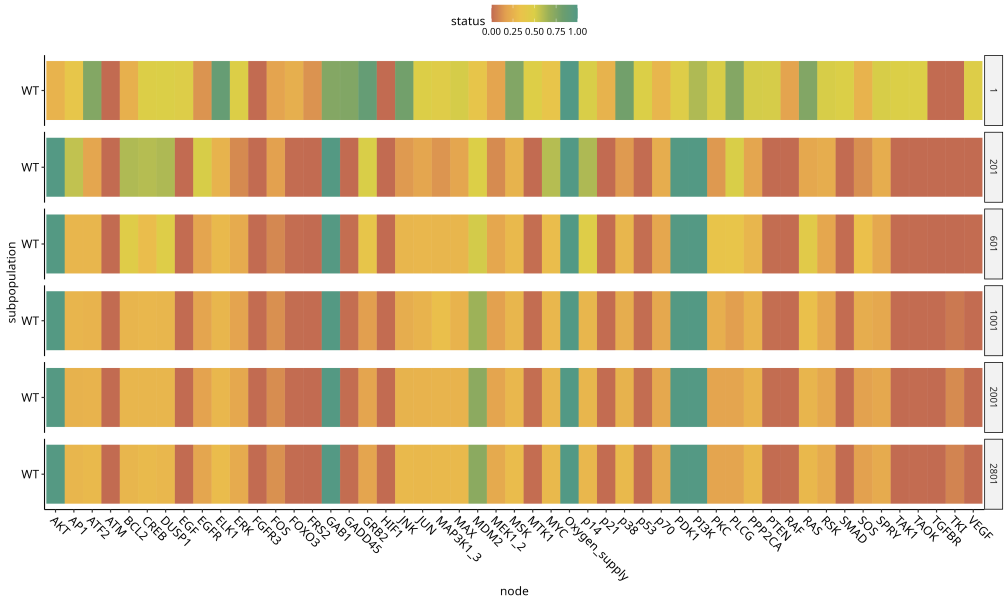


(b)

Figure 4.24: Diffusion and intake of a timed drug treatment. A simulation with drug treatment was run with 100 WT cells. The drug, which blocks the EGFR node, became active at t=200, and inactive at t=2000. The numbers above each figure are the time points. (a) shows the diffusion of the TKI drug across the 2D space in a similar manner to figure 4.11. (b) displays every cell in the simulation coloured by their activation of the TKI node. An active TKI node means that that cell has successfully absorbed the drug and it is causing an effect in their network.

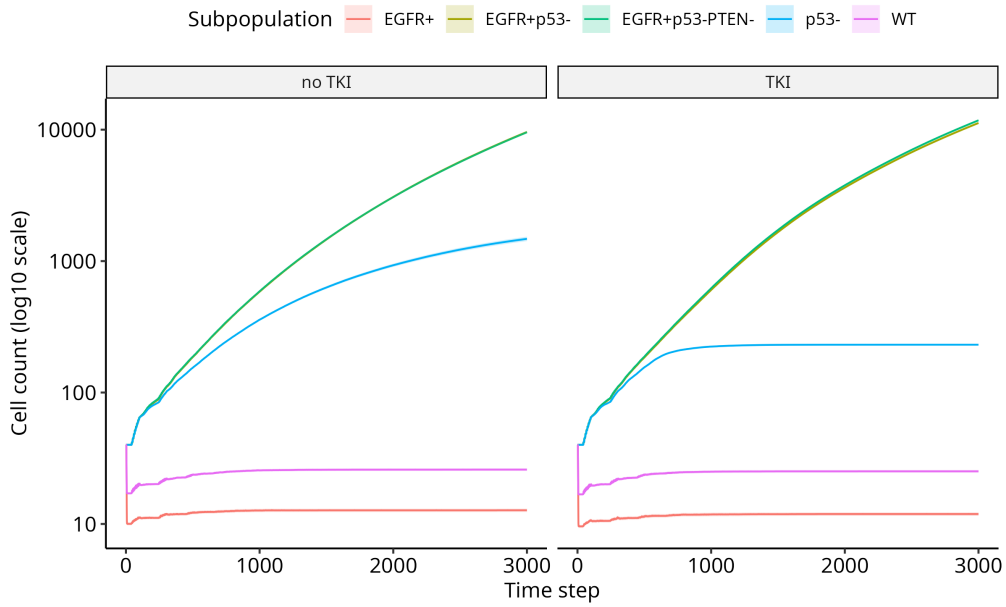


(a)

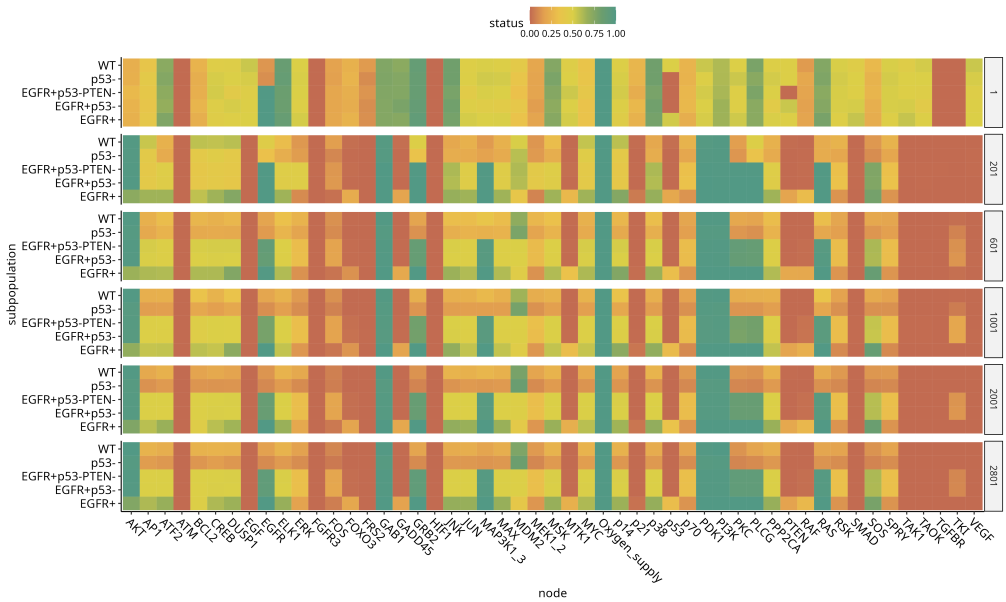


(b)

Figure 4.25: Effect of a EGFR-blocking drug on cell proliferation. Results of a simulation run with the same parameters as for figure 4.24. The figures are the average of 100 replicates, with an initial population of 100 cells, run for 3000 time steps, with drug treatment starting at t=200 and ending at t=2000. (b) shows the activation status of nodes in the network of each cell, averaged by time point, during the treatment simulation.



(a)



(b)

Figure 4.26: Effect of a EGFR-blocking drug on competition. Results of a simulation run with WT, EGFR+, p53- EGFR+p53- and EGFR+p53-PTEN- competing subpopulations. The figures are the average of 100 replicates, run for 3000 time steps, with drug treatment starting at t=200 and ending at t=2000. (a) and (b) as in figure 4.25

as gefitinib, with similar diffusion parameters to such a drug. In figure 4.24 a one can see the diffusion of the TKI in a simulation where it starts being supplied to the tumour at t=200 and stopped at t=2000. The dark circle in the centre of the simulation in later time steps is evidence of cells on the outside of the tumour consuming the drug — depending on the concentration at the boundary and the treatment duration, it is possible that only the outer cells will present drug intake. The penetration of the drug in the tumour during this specific simulation is represented in figure 4.24 b , where drug consuming cells are coloured in yellow.

Figure 4.25 demonstrates the effect of the TKI on the growth of a wild type population (figure 4.25 a). After the end of the treatment (t=2000) the cells do not accelerate their growth, which shows that the drug effect lasts for a while after their supply to the simulation (figure 4.24 a). The specific effect of the drug on the cell's network is shown in figure 4.25 b , where from t=200 EGFR and its downstream (GRB2, RAS, SOS, ERK, ...) are inhibited. The inhibition of EGFR could be forcing the cell into a series of network states from which reaching a proliferation attractor takes longer than for an undrugged cell. Or they could be going into apoptosis before they can reach the proliferation attractor, in which case they can be replaced by new cells that have never been exposed to the drug.

Furthermore, figure 4.26 contains the results of EGFR TKI treatment on several subpopulations growing in competition. It can be seen that EGFR+ cells are predicted to be quite resistant to this treatment. Even though this simulation was run with a synchronous network and therefore the activation mutation of EGFR is total (that is, EGFR is permanently activated in those cells), that mutation does not override the action of the drug. In other words, if the drug reaches an EGFR+ cell, EGFR will be deactivated in that cell as long as the concentration of the drug is above the activation threshold. That the activation of EGFR in mutant cells is not at its maximum during drug treatment can be surmised from figure 4.26 b .

Still, this is one of the main flaws in how drugs are being modelled in the current *microCJ*. Because I designed them to only activate their input node in the network for one step —as long as the drug concentration is above threshold— any time the

drug is momentarily not present in the cell's voxel —because it has been consumed by other cells in that and neighbouring clusters, and it will take more than one time step to the drug concentration to go over the threshold— the input node will be turned off, and since the cells have a EGFR+ mutation, they are predicted to proliferate. To more closely model reality, drugs should have a more lasting effect, and EGFR+ cells should be specially sensitive to EGFR inhibition, but also the network I am using may not be complex enough for this. Intake and efflux rates could be modelled, as well as the degradation time for the drug. It would also be worthwhile to look into how EGFR+ cells come to depend on that pathway, so that they are more sensitive to it than other cells.

Another issue with the current simulation is that from figure 4.24 b , it seems that the drug is only reaching the outermost cells. But it cannot be known for certain. Since, as mentioned, the effect of the drug only lasts for one time step, it could be reaching all cells but only showing up in the outer ones because those are the only ones exposed to the drug at every time step. The growth curves show a considerable effect of the drug, so it would seem that most of the cells are receiving it, but this is something else that should be improved in subsequent iterations of *microCJ*.

That said, the drug does cause an effect on the p53- population, which goes from moderately growing in the absence to the drug to almost being killed off in its presence. The lack of a big p53- population plays to the advantage of the EGFR+ cells, that take advantage of the space left empty and reach a bigger growth that in the absence of the drug.

Finally, the cells can gain resistance to a given drug. It can be specified in the network file that a drug node can possess a certain probability to acquire resistance to the drug, modelled as a probability of the node staying off despite the drug reaching the activation threshold. This probability can become larger in successive time steps, thus modelling mutations that decrease the affinity of the drug to the binding site, as it is the case with EGFR TKIs, for example[340,341].

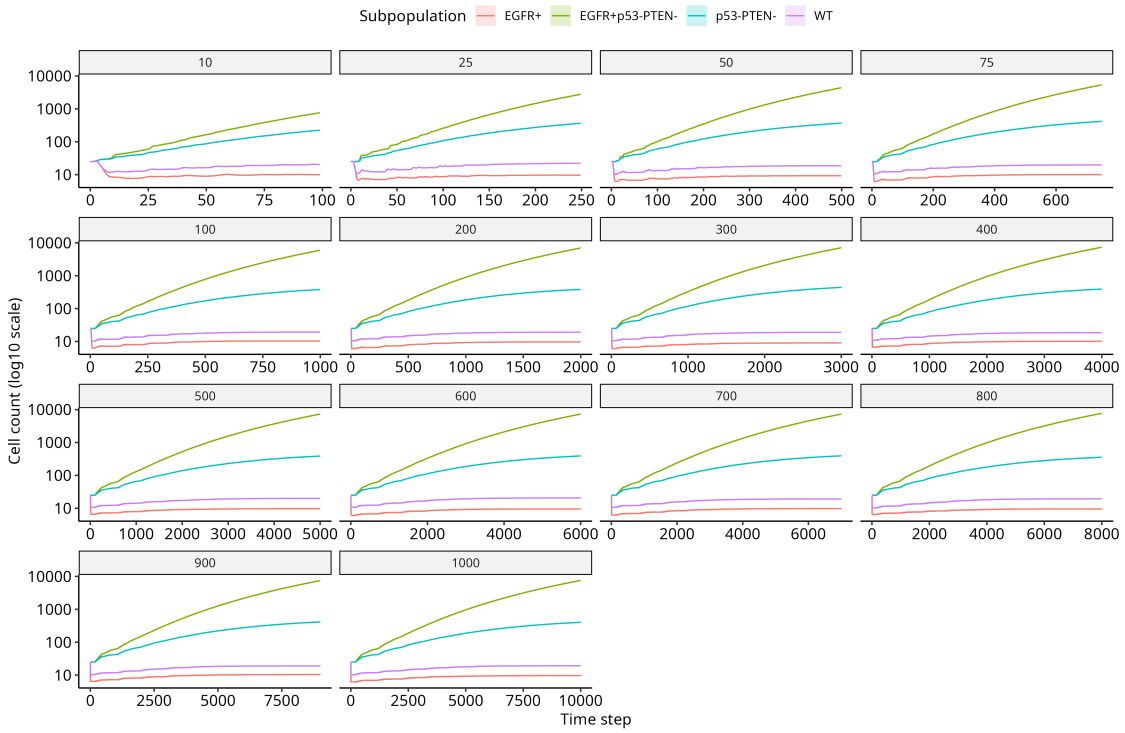


Figure 4.27: Decision window sensitivity analysis. Fourteen competition simulations were performed, with the same parameters as figure 4.5, but each of them with a different decision window ranging from 10 to 1000. The label above each growth curve plot denotes the length of the decision window. Each of the plots shows the means of 100 replicates. Each of the simulations were run for 10 decision window cycles, this is, 10 * decision window length.

4.2.7 Sensitivity analyses

4.2.7.1 Instant growth arrest

In this section I will test and perform sensitivity analysis on several mechanistic and technical parameters to assess the impact of different values in the behaviour of the model: how robust it is to changes in these parameters. This identifies the essential parameters, this is, the ones most influence the system and should be calibrated with biological data to create a credible model.

4.2.7.2 Decision window

First I test the decision window. As previously explained, the decision window is the length in time steps of the cycle at the end of which proliferation and growth arrest decisions can be executed. This wait, as opposed to apoptosis and necrosis,

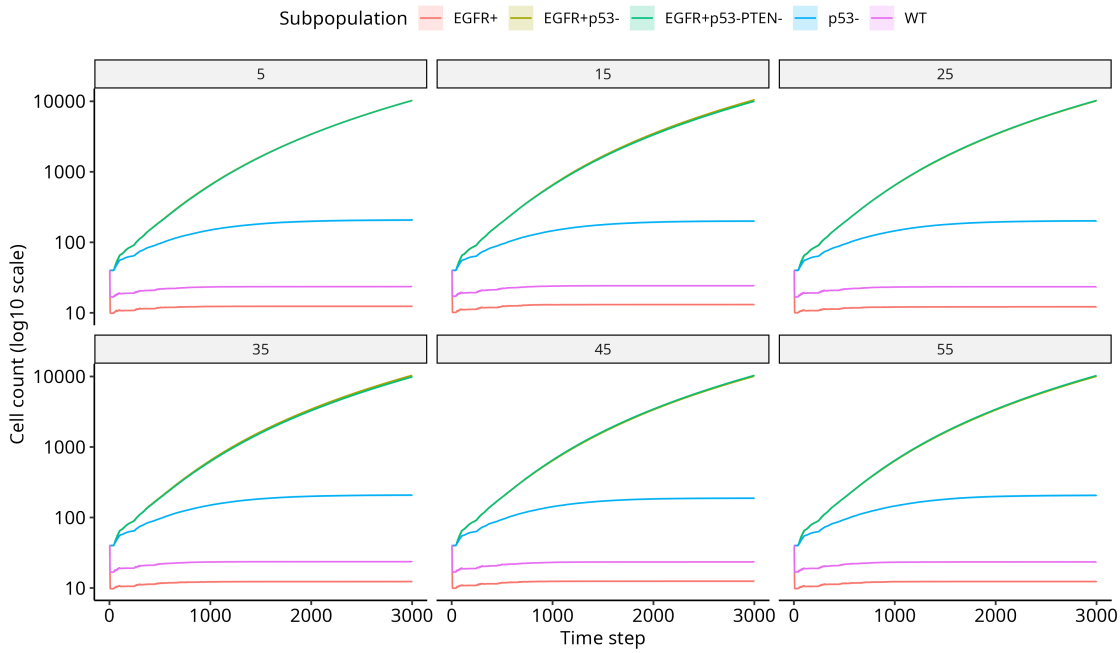


Figure 4.28: Growth arrest duration sensitivity analysis. Six competition simulations were performed, with the same parameters as figure 4.5, but each of them with a different growth arrest duration ranging from 5 - 55. Each of the plots shows the means of 100 replicates, running for 3000 time steps.

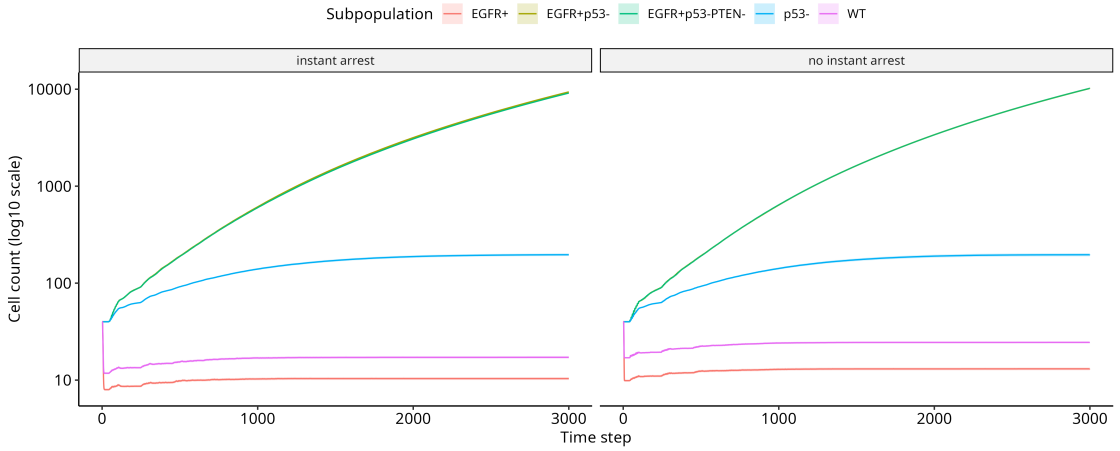
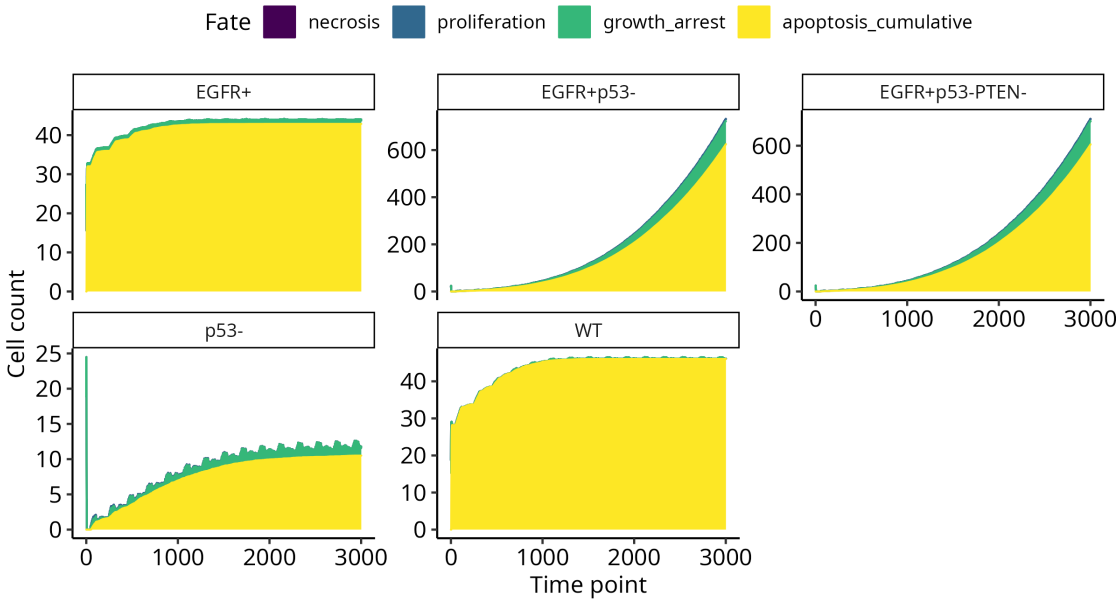
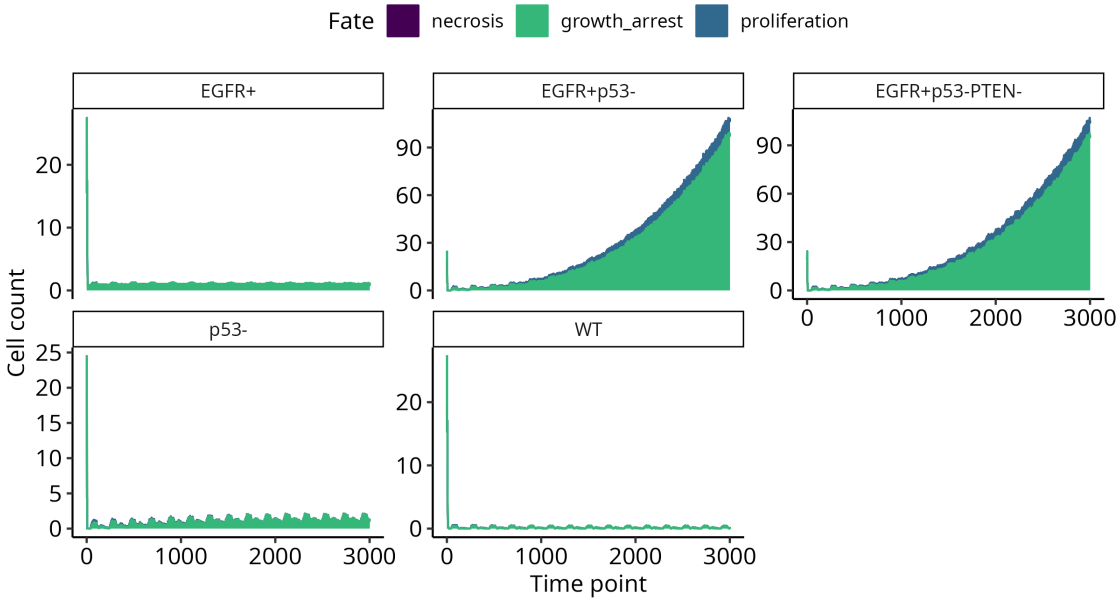


Figure 4.29: Instant growth arrest comparison. Competition simulations, with the same parameters as figure 4.5, but one of them has instant arrest enabled and the other one does not. Each of the plots shows the means of 100 replicates, running for 3000 time steps.

that are executed immediately, mainly represents the amount of time needed for the cell to grow and progress through the cell cycle phases before it can start the mitosis. But one also has to bear in mind the minimal amount of time needed to reach an



(a)



(b)

Figure 4.30: Instant growth arrest relationship with overall fate activation. For a explanation of the plot, see figure 4.8. Data for this figure came from the same simulation as figure 4.29.

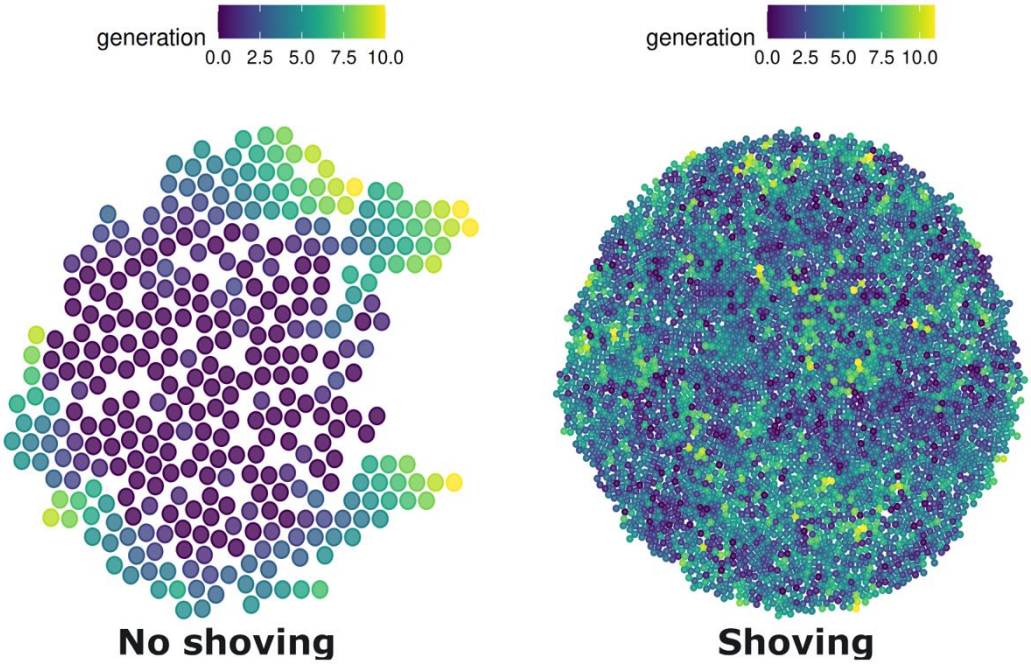


Figure 4.31: Effects of the "shoving" parameter on the growth of a tumour. The tumour on the left was run with the parameter shoving set to off, and the one on the right had it set to on. Shoving means that if a proliferating cell cannot find an empty space next to it, it will not go into growth arrest (the default behaviour) but will push all other cells around it to make enough space for the new cell.

attractor in the Boolean network inside each cell — as discussed before, attractors roughly model the phenotype of the network, the cell fate decisions in this case. The decision window should not be smaller than this minimal amount, otherwise the phenotype of the model cell would be a function almost exclusively of its initial state (which is random) and not of its relationship with its microenvironment, the other cells, and its genotype (the "mutations" of the Boolean network). This would be undesirable. For the unperturbed, inputs-off synchronous network (figures 3.4 and 3.5), the average number of transitions to reach an attractor is 11.55, the maximum 30. Figure 4.27 shows the result of these analyses. The general shape the growth curves is similar across all the different time windows, with exponential growth for the two most aggressive subpopulations already observed around decision window = 25. Around decision window = 200 steps the size of each subpopulation at the end of the simulation begins to be consistent. This is the point I have chosen for all the other simulations, which is well over the maximum transitions

needed for an attractor to be reached. Note that the length of each simulation is different, each of them runs for 10 times the decision window length. As previously explained each cell starts the simulation at a different point of their cycle, so they do not divide in synchronicity. Still, they roughly go through the same amount of proliferation/growth arrest cycles, as it can be seen in the similar end subpopulation sizes from around 100-200, even though for decision window 100 the simulation runs for 1000 steps and for decision window 1000 it runs for 10,000.

4.2.7.3 Growth arrest duration

The growth arrest duration is an arbitrary parameter, and in figure 4.28 it can be seen that it does not have much weight on the overall behaviour of the model. This is certainly due to the fact that most growth arrest network activation is ignored. As shown in section 3.1.1, most of the time growth arrest happens in the same attractor and state than apoptosis. In *microCJ*, when these two fate nodes are activated at the same time, the cell goes into apoptosis, and therefore the growth arrest duration does not matter. Apart from network decision, the other way for a cell to go into growth arrest is if it tries to proliferate but there is no adjacent space (and shoving is not enabled). But a cell that has no space around it is unlikely to ever have it, so again the growth arrest duration does not matter, if this MAPK network is used. However, there could be exceptions, like a population with certain mutations that protect it from going into apoptosis too often, but can go into growth arrest, like the one in figure 3.10.

Another parameter that needed to be tested is if the growth arrest decision should happen immediately after the node is activated in the network or if on the contrary should wait for the cell cycle (the decision window) to end. As arrest is a possible outcome of the cell cycle, it makes sense for it to be modelled as part as the same system as proliferation. Still, cells that go into arrest stop the cell cycle sooner than cells that go all the way up to mitosis. I compared both alternatives (figure 4.29) and it makes little difference to the overall shape and order of the growth curves. Still, the simulation with instant arrest achieved a lower end subpopulation size, likely

due to those few attractors that activate apoptosis and growth arrest at different states of their cycle, so if the state with growth arrest is reached before the apoptosis one, that will be the outcome, or even if growth arrest is transiently activated outside of an attractor. Those cases would not cause an effect during non instant growth arrest simulation, because they would not be taken into account unless they co-occur with the end of a cell decision cycle; but they would with instant growth arrest. And if one compares figures 4.8 and 4.30, one can see a small increase in growth arrest activation in the instant arrest simulation. In any case, this did not seem like an essential parameter, and therefore I decided to use the non instant arrest option, due to the biological significance and a lower computational load.

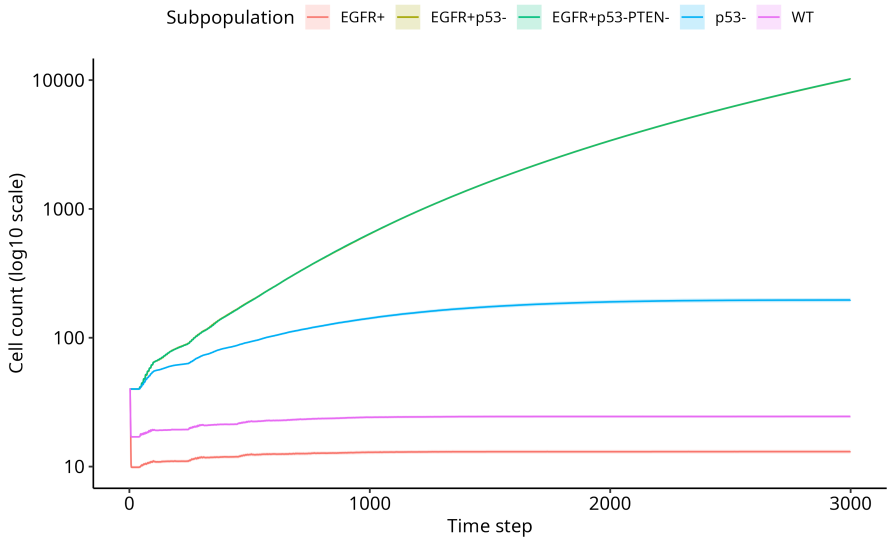
4.2.7.4 Shoving

The difference between a tumour in which cells are densely packed and do not have much space to grow, and a more lax one can be simulated using the "shoving" parameter. Most of the simulations in this thesis were run with this parameter set to off, which means that when a cell cannot find a space next to it to place its daughter cell, it will not produce one and go into growth arrest instead (simulating contact inhibition). Alternatively, if shoving is set to on, a cell that founds itself with not empty spaces to proliferate will instead physically push all other cells to make enough space next to it. This "push" is transmitted to all other cells in the simulation, which will move just enough so that no two cells are overlapping. The difference this makes in the overall tumour progression can be observed in figure 4.31, which presents the cells coloured by generation. The tumour on the left was simulated with no shoving, and consequently newer cells are on the outside, since all space in the core of the tumour is occupied. It can be seen that there are gaps in between cells on the inside — this is due to inefficient random packing of the cells, which can leave these empty spaces but that are not big enough to contain a new cell. On the contrary, newer cells can be anywhere in the shoving simulation (right tumour), and they are more densely packed as cells can be pushed to make empty

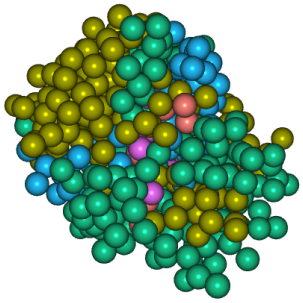
spaces big enough for new cells. Since no cell will go into growth arrest because of contact inhibition, the tumour on the right is much bigger than the shoving=off one.

4.2.7.5 Cell placement method

Another relevant cell physical model parameter is the placement method. *microCJ* supports both 3D and 2D simulations, and for each of them there are several options, as previously explained. Figures 4.32 to 4.36 present a comparison between all the different methods. The growth curves allow us to see that generally, 2D simulations achieve a lower end subpopulation size than 3D. This seems logical when taking into account that in 3D simulations a cell has a higher maximum number of neighbours, and therefore will divide more times before running out of space and going into growth arrest. Similarly, there is more growth in 2D square (figure 4.36 a) than in 2D hexagonal (figure 4.34 a), again due to a higher number of possible neighbours. And any lattice based placement method has a higher growth rate than their 2D or 3D random counterpart, because the randomness provokes a more inefficient packing of cells. The age plots show the way tumours place the new cells according to the selected strategy. This is just a function of the model assumptions, and cannot be used to extract any biological conclusions, since *microCJ* does not include cell-cell mechanical interactions. At any rate, there is no reason to categorically prefer one placement strategy to another, as the growth curve shapes and relative success of each subpopulation is the same for every one of them. Instead, it can be decided taking into account the acceptable computational load and requirements of every experiment: random placement is more computationally expensive than hexagonal or square (and square more than hexagonal), 3D placement is more realistic, 2D placement allows for easy visualisation of tumour architecture and is computationally lighter.

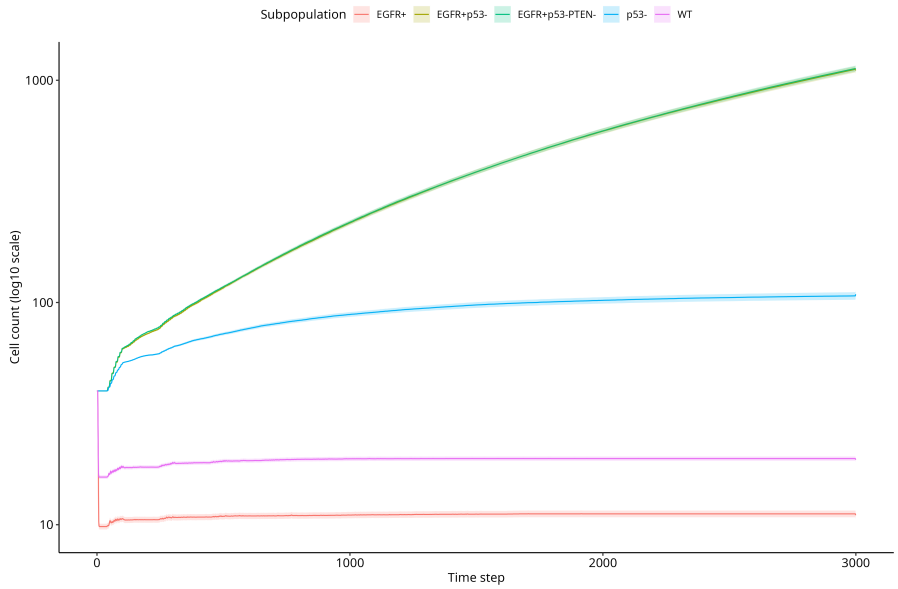


(a)

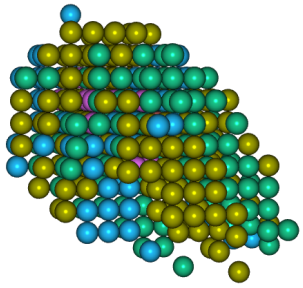


(b)

Figure 4.32: 3D random placement. The 3d random placement strategy. (a) shows the growth curves of the simulation from figure 4.5, which was run using this strategy. (b) shows a screenshot of the graphical representation in the *microCJ* user interface.

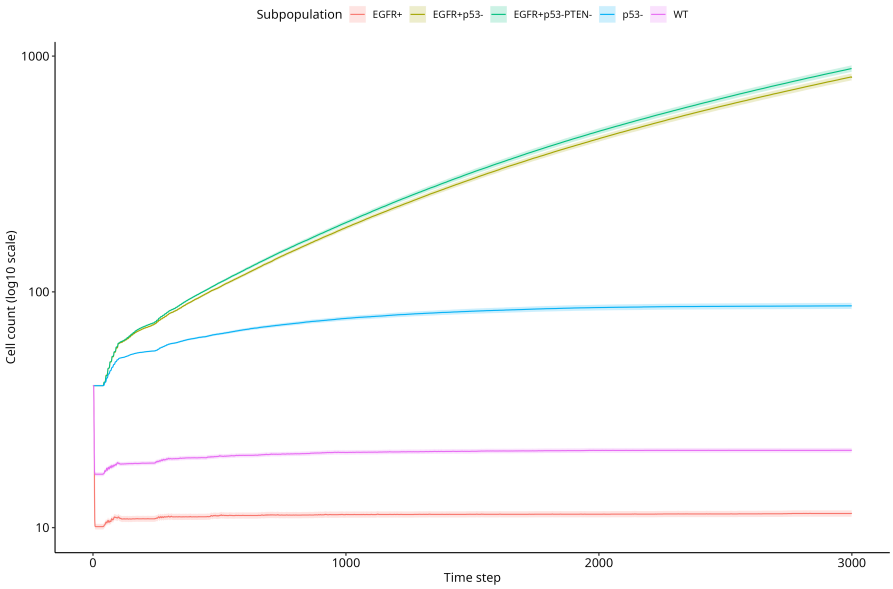


(a)

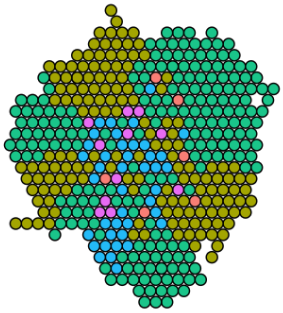


(b)

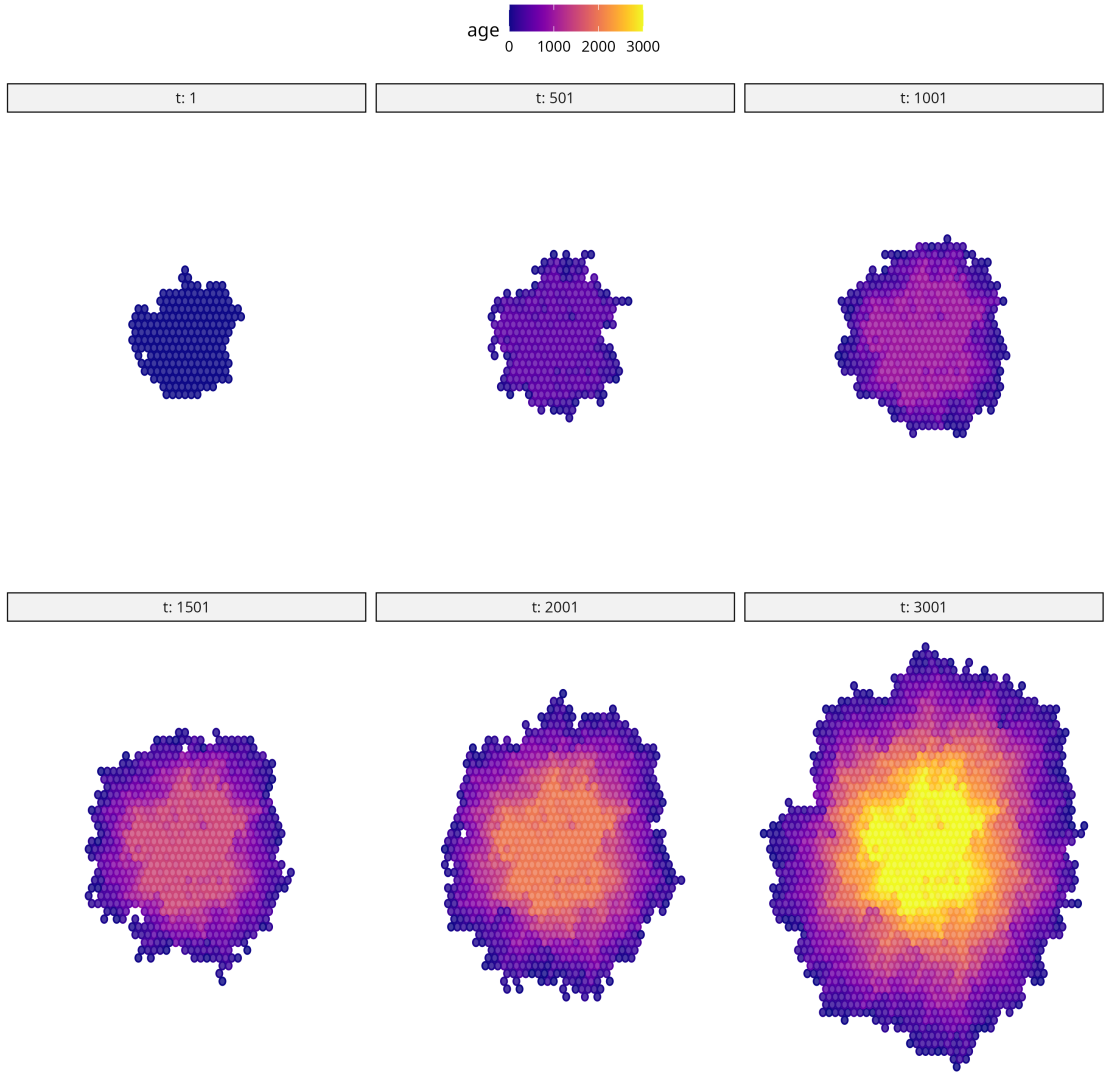
Figure 4.33: 3D square placement. (a) shows the growth curves of a competition simulation with the same parameters as figure 4.5, but using this placement strategy instead. (b) shows a screenshot of the graphical representation in the *microCJ* user interface.


Subpopulation
EGFR+
EGFR+p53-
EGFR+p53-PTEN-
p53-
WT
1000
100
10
Cell count (log10 scale)
0
1000
2000
3000
Time step

(a)

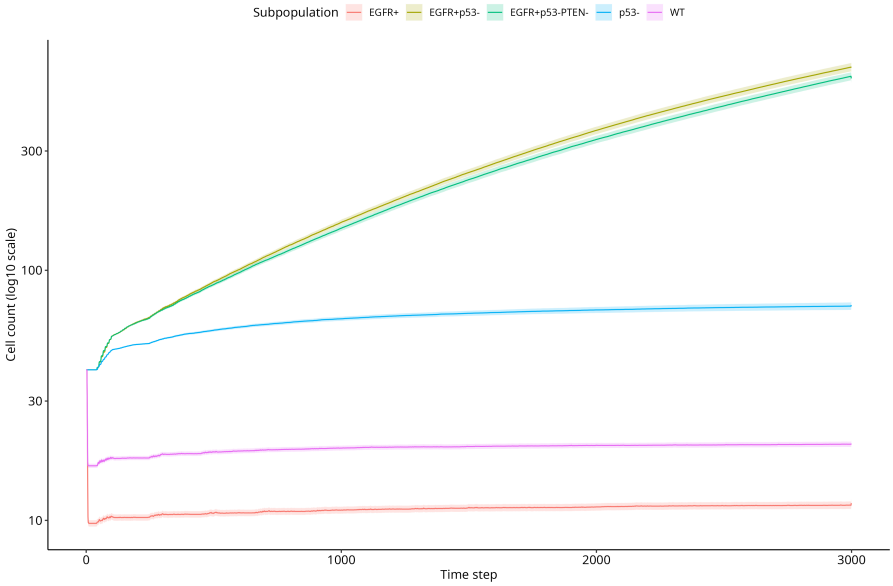


(b)

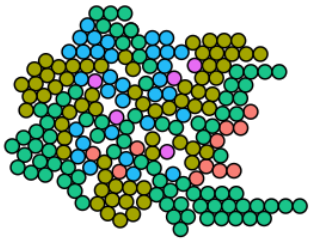


(c)

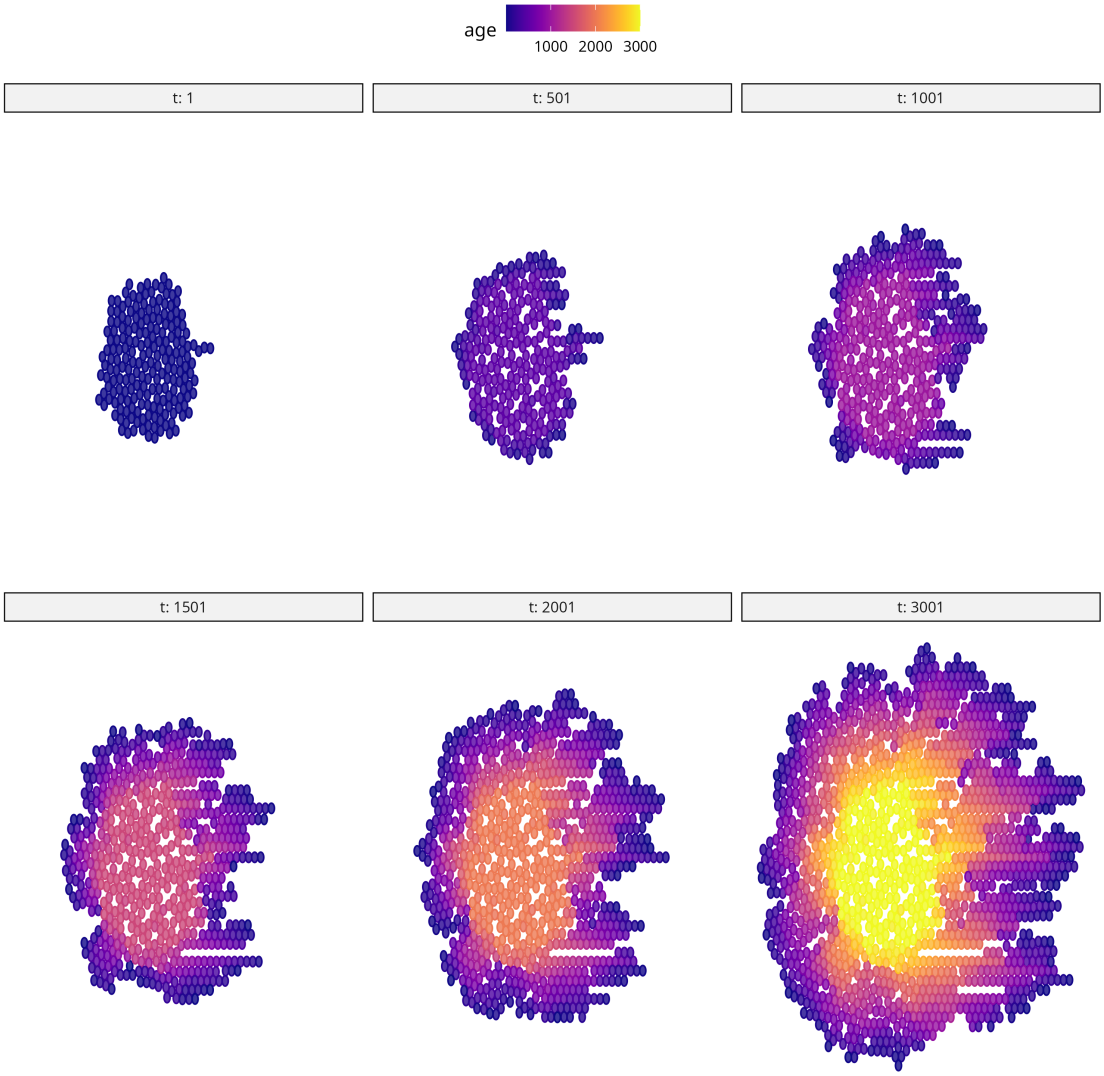
Figure 4.34: 2D hexagonal placement. (a) shows the growth curves of a competition simulation with the same parameters as figure 4.5, but using this placement strategy instead. (b) shows a screenshot of the graphical representation in the *microCJ* user interface. (c) plots the cells at the end of the simulation coloured by their age, this is, how many time steps have passed since they were created.


Subpopulation
EGFR+
EGFR+p53-
EGFR+p53-PTEN-
p53-
WT
Cell count (log10 scale)
300
100
30
10
0
1000
2000
3000
Time step

(a)

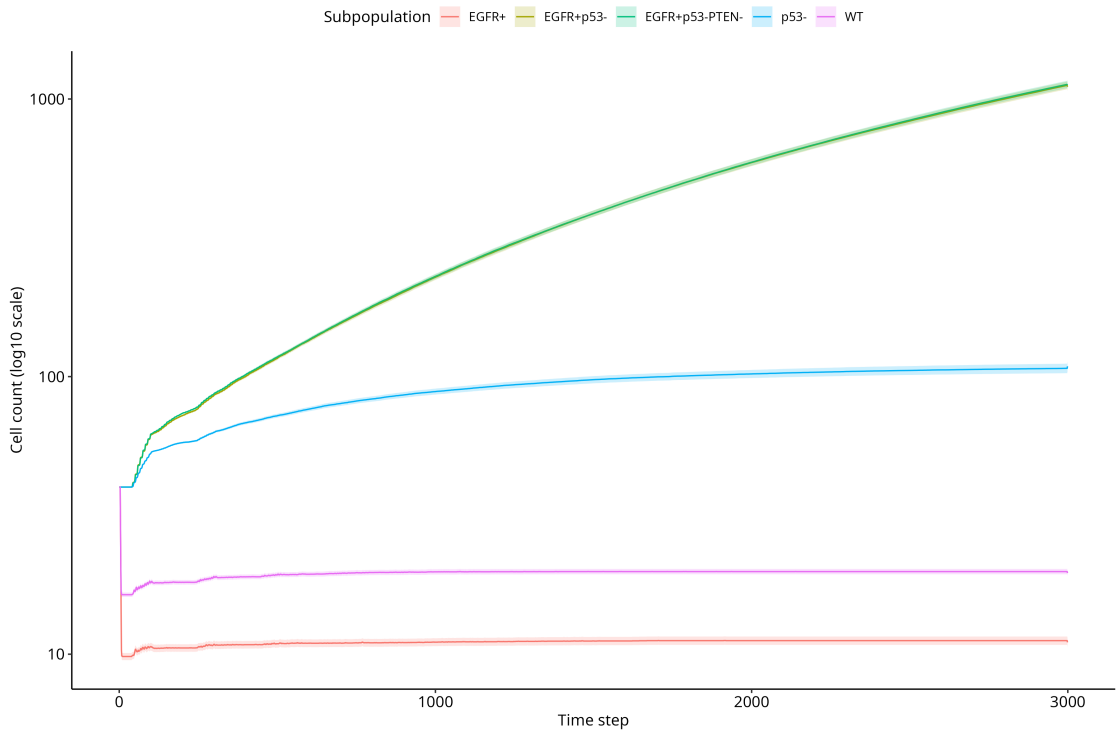


(b)

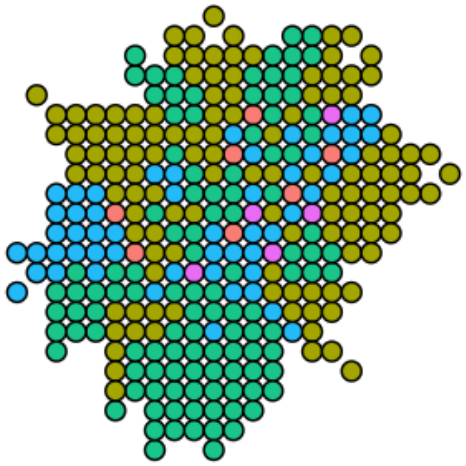


(c)

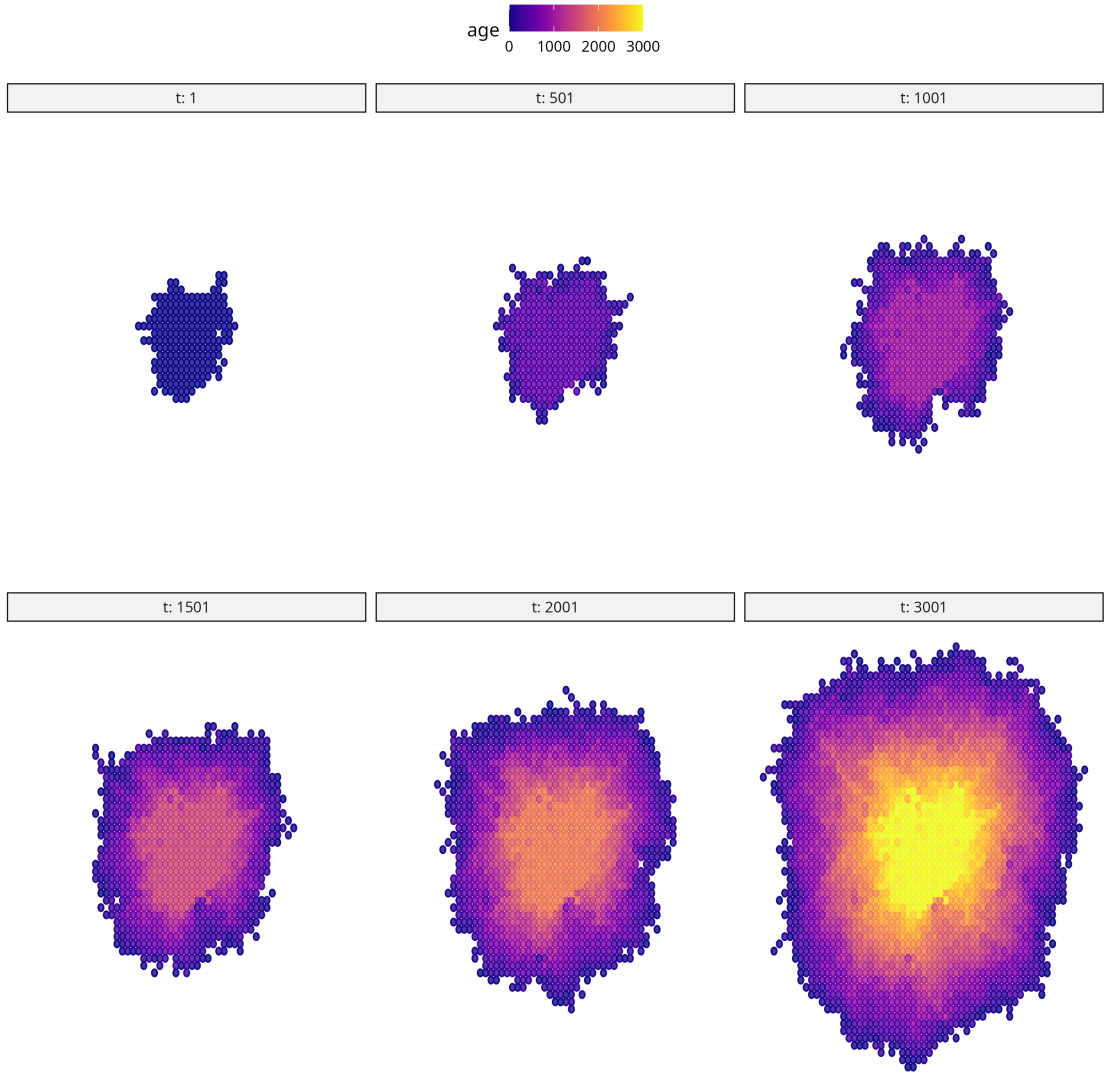
Figure 4.35: 2D random placement. The 2d random placement strategy. (a) shows the growth curves of a competition simulation with the same parameters as figure 4.5, but using this placement strategy instead. (b) shows a screenshot of the graphical representation in the *microCJ* user interface. (c) plots the cells at the end of the simulation coloured by their age, this is, how many time steps have passed since they were created.


Subpopulation
EGFR+
EGFR+p53-
EGFR+p53-PTEN-
p53-
WT
1000
100
10
Cell count (log10 scale)
0
1000
2000
3000
Time step

(a)



(b)



(c)

Figure 4.36: 2D square placement. The 2d square placement strategy. (a) shows the growth curves of a competition simulation with the same parameters as figure 4.5, but using this placement strategy instead. (b) shows a screenshot of the graphical representation in the *microCJ* user interface. (c) plots the cells at the end of the simulation coloured by their age, this is, how many time steps have passed since they were created.

4.3 Conclusions

In this chapter, I have:

- Described the *microCJ* model and its features. *microCJ* is a multiscale model of tumour growth with a large variety of features and parameters, which allows researchers to model an array of different scientific questions.
- Shown that the model, while working on the basis of simple cell agents with a simplified gene regulatory network, displays collective behaviours that are coherent with biological knowledge, especially through the use of probabilistic Boolean networks to simulate different levels of gene and protein activity, and mutation acquisition to simulate tumour evolution.
- Analysed the model parameters, to test its robustness and the role they play in the overall behaviour of the tumour.

The mechanistic, bottom-up design of *microCJ* provides capability for studying both the causes of behaviour at the single cell level and the emergent properties of these behaviours. The combination of agent-based modelling and the probabilistic Boolean network setup plays an essential role here: it allows hypotheses to be tested in a controlled environment, divided in discrete steps of time and space.

Multi-cellular dynamics are astonishingly complex, and models give us a way to dissect this complexity. By building up on very simple components, but treating them as independent agents, *microCJ* can simulate the evolution of these complex multicellular dynamics, the interaction and intracompetition in a heterogenous population, and the relationship between individual cell genotypes and the tumour phenotype. Indeed, the agent-based approach is specially interesting in the context of the single-cell sequencing data explosion the scientific world is experiencing at the moment, where understanding in depth this relationship between genotype and collective phenotype is more relevant than ever. On that note, one of *microCJ* features is the ability to import single-cell and spatial transcriptomics data, preserving cell by cell information, through the use of *Ingres* (described in chapter 3).

Cells in *microCJ* are not identical and monolithic: they can display different sets of mutations, and thanks to the probabilistic setting, they can present continuously different levels of expression in any gene. They can also evolve and gain new mutations, which makes enables *microCJ* to model an ever changing and dynamic tumour, closer to biological reality. This allows researchers to cover many possible scenarios that other kinds of models would not be able to reach. For example, studying the properties of a gene regulatory network in isolation, as I have done in chapter 3, will inform us of the possible outcomes of the modelled pathway, but will not give us any insight into which of them are more likely to occur *in vivo*, which will give cells a competitive advantage, how each of these outcomes will affect the whole tumour, what effect does the spatial context have on them, and so on.

Furthermore, *microCJ* not only takes into account the cell genotype, but also important aspects of the microenvironment such as oxygen diffusion — response to hypoxia is one of the most important determinants of tumour evolution — growth factors, and the physical interaction between cells, who must compete for space to grow in and to not become cut off from the oxygen supply.

There are several shortcomings of the modelling framework presented here. First, despite being built on knowledge derived from experimental biology, the model predictions remain to be experimentally validated, an essential step in the development of any modelling tool. This will be the most important priority in the future of this project. Second, the diffusion modelling has some flaws, namely the conflation of the diffusion and intracellular time scales, which is not realistic. This should be improved in subsequent versions of *microCJ*, so that the network simulation, the intercellular simulation and the diffusion model operate on different time scales. Third, the modelling of drugs in *microCJ* is flawed, and it needs to be revised, so that drugs have a more lasting effect, and their diffusion work in a more realistic manner. Finally, the modelling framework should be tested for computational complexity, how simulation time increases with increasing network complexity, diffusion domain size, or number of cells.

In this chapter I have illustrated a range of potential applications, such as the study of how specific mutations occurring in a subpopulation of cells can affect the collective behaviour of the whole system. Among those mutations, I have pointed at at some of the most frequently observed across several cancer types, such as p53 and PTEN deletion or EGFR over-expression. *microCJ* shows that cells carrying two or more of these mutations have an advantage over ones just having one of them. This advantage is enhanced when growing under adverse conditions such as a hypoxic environment. Moreover, when the model is left to evolve on its own it ends up selecting the same mutations, and others that are also frequently observed in tumours, such as ERK or AKT over-expression.

The addition of probabilistic Boolean networks opens up a vast array of possibilities. By modelling mutations and differential expression in a more detailed manner than just total deactivation or over-expression, *microCJ* addresses one of the main limitations of previous models, while preserving the computational economy and intuitiveness of Boolean networks. With PBNs, *microCJ* can model small differences between subpopulations in a granular way, providing the tools for simulation of clonal selection and gradual evolution.

All in all, *microCJ* is not only an essential part of my DPhil project, but it has the potential to become an invaluable tool to aid researchers to test hypotheses and generate new ones in a controlled, knowledge-driven framework, and to make mechanistic sense of highly multi-dimensional transcriptomics data.

4.4 Methods

4.4.1 Implementation

I am the sole author of all the *microCJ* codebase, including the graphical capabilities. *microCJ* is coded in Java 16, and for the graphical rendering it makes use of the JavaFX framework. As an object-oriented language Java lends itself well to the agent-based abstraction, since every agent is coded as a separate software object with the corresponding value for each of its parameters. *microCJ* is

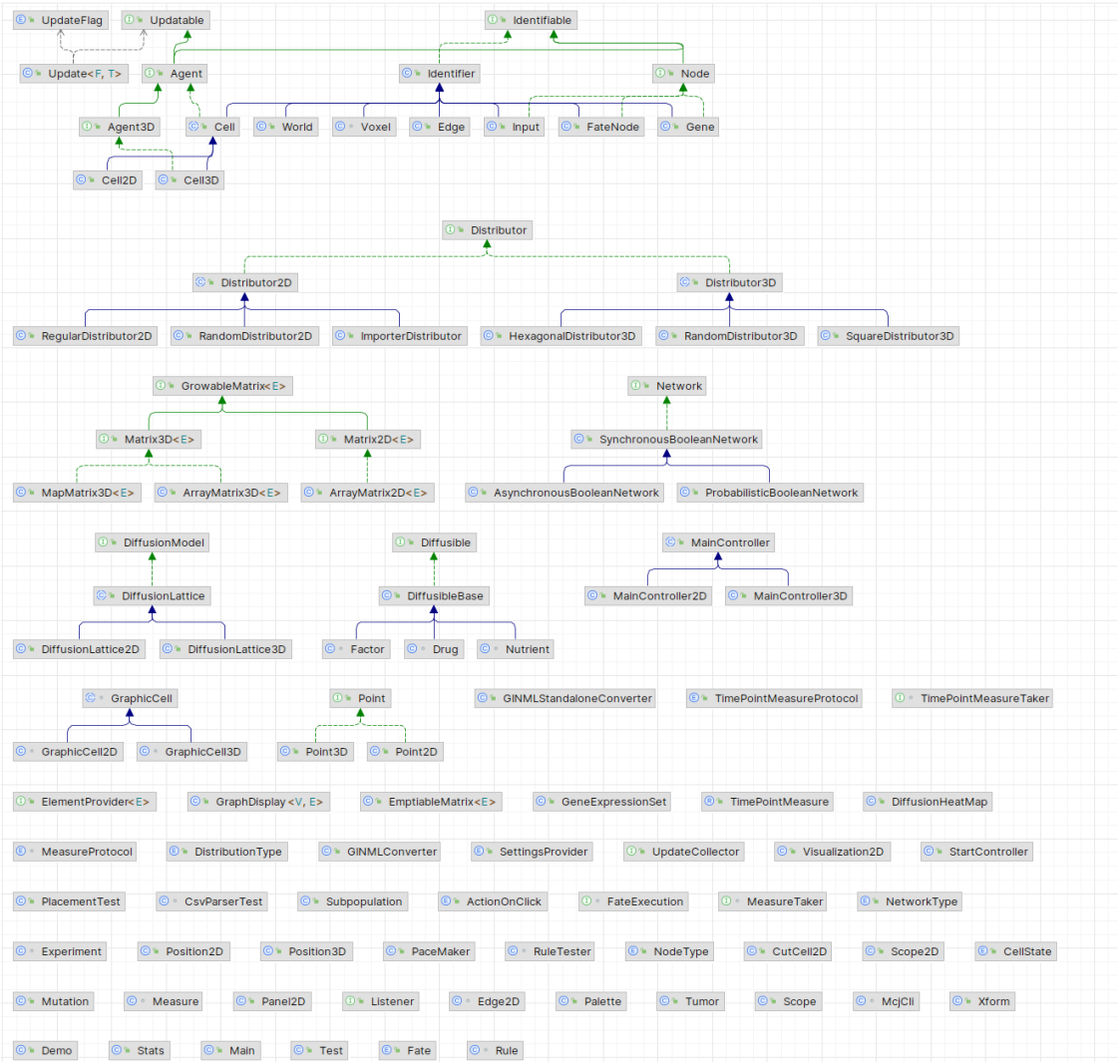
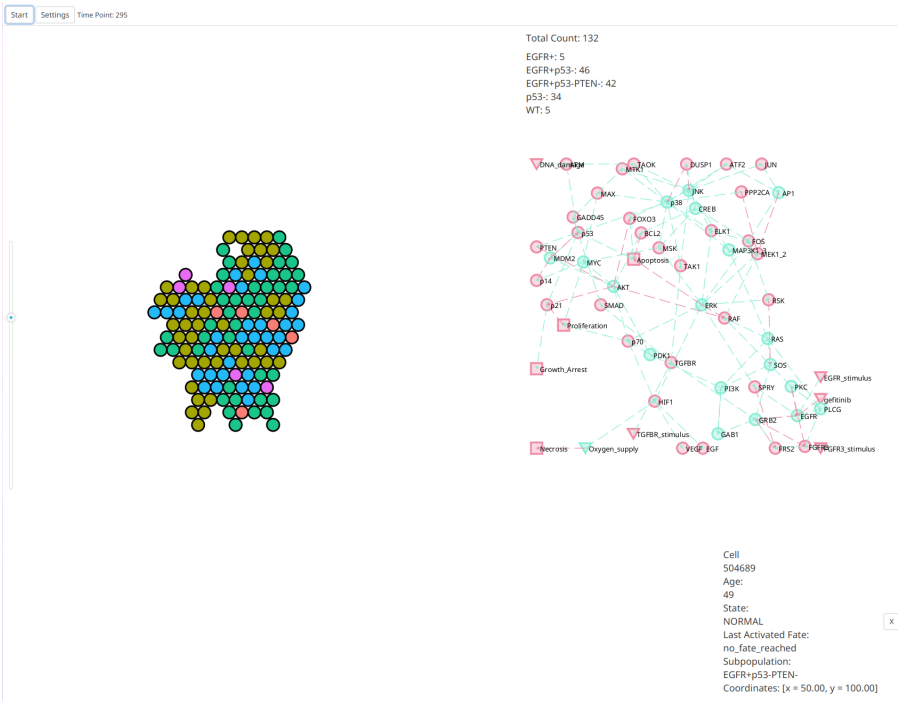


Figure 4.37: Overview of *microCJ* codebase structure. The green letter I signifies that the element is a Java Interface, the blue C means a Java Class. E means an Enumeration. Solid green arrows represent an interface extending another interface, the one pointed by the arrow. Solid blue lines represent classes extending another. Dashed green lines represent classes implementing an interface.

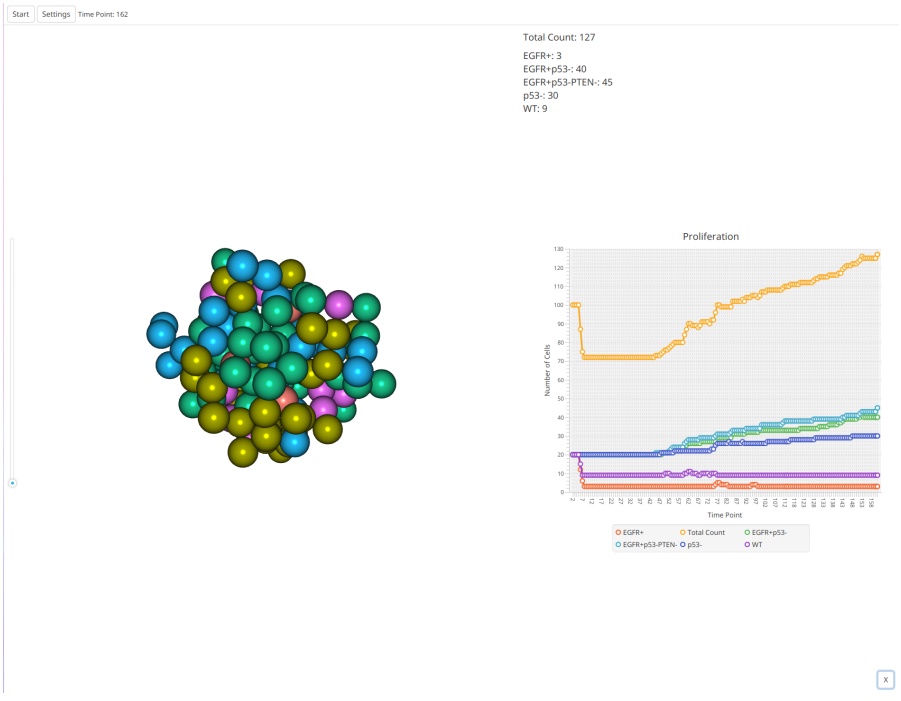
open source, released with a AGPL-3.0 license, and its source code can be consulted on Github (`https://github.com/CBigOxf/microCJ`.

microCJ implementation is modular, every subsystem is programmed as a separate module with just a few channels of data transfer between itself and the other modules figure 4.37. This allows researchers to extend and modify the specific functionality they are interesting in, without a need to know the whole codebase.

microCJ is fast (figure 4.39) and makes use of multi-core programming, which



(a)



(b)

Figure 4.38: A view of *microCJ* GUI.

allows users to take advantage of modern, powerful computers to speed up the simulations.

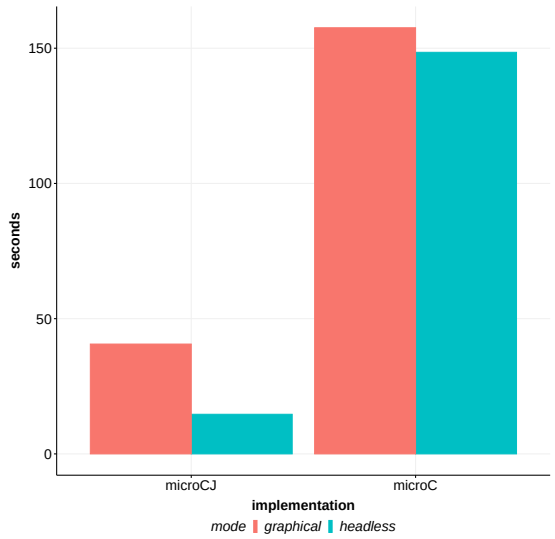


Figure 4.39: Performance improvement in microCJ. Compared performance in microC and microCJ, measured as the time (in seconds) it takes to run 2000 time steps of a simulation with an initial cell count of 100, with the default synchronous MAPK network, without any other perturbations, in normoxic conditions, both without producing a graphical representation ("headless", blue) and with it (red). Simulations were run in a computer with the following specifications: i5-8600K, 32GB RAM DDR4, GTX 1060 6GB VRAM.

4.4.2 CLI and GUI

microCJ includes a command-line interface (CLI) and a graphical user interface (GUI) to allow an user to control it. *microCJ* could also be imported as a standalone library and be connected to custom interfaces.

The GUI is primarily meant for preliminar exploration of simulations, running a single simulation and showing results such as the tumour architecture, cell Boolean networks, growth curves and diffusion heatmaps in real time (figure 4.38).

On the other hand, the CLI allows the user to run hundreds of replicate simulations in parallel, and to perform sensitivity analysis for any parameter, by expressing it as a range instead of single value in the settings file. A list of the parameters the CLI takes can be seen in its help message:

```
microcj [-hV] <settings> <iterations> <steps>

A 3D, multilevel agent-based tumour modelling framework.
    <settings>    The xml file containing the desired settings.
    <iterations>  Number of simulations to be run.
    <steps>       Number of steps each simulation will run for.
```

```
-h, --help          Show this help message and exit.
-V, --version       Print version information and exit.
```

The user can specify different setting files for different simulations, so that many different configurations can be tested at once. Due to the computational expense of running several replicates at once, especially if the number of replicates is bigger than a few dozens, the CLI is particularly suited to be run in a server or cluster configuration.

After each simulation has run for the required number of steps, the CLI produces all the output files as specified in the `stats` and `timepoint-stats` groups of settings.

The GUI also creates outputs files, but these are updated at every time step, unlike the CLI which waits until the end of the simulation. This is because GUI simulations do not have a defined end point. Therefore they can be stopped at any time without data loss.

4.4.3 Settings

microCJ accepts a single xml setting file, where every parameter of the simulation can be configured. Table 4.1 shows a full description of each one of them.

Section	Setting	Value	Explanation
general	starting-cells	A positive integer	Total number of cells at the beginning of the simulation
general	cell-radius	A positive integer	Cell radius in µm
general	graph-file	A string	The name of the .graphml file containing the Boolean network to be used in the simulation
general	decision-window	A positive integer	The length of the decision cycle

Section	Setting	Value	Explanation
general	arrest-time	0 or a positive integer	The duration of growth arrest. Setting it to 0 disables growth arrest altogether.
general	instant-arrest	A Boolean	Whether growth arrest is instant, otherwise it is part of the decision cycle.
general	initial-states	A Boolean or a number between 0 and 1	If false, all nodes start the simulation inactive, and vice versa if true. If its a number, that denote the probability of each node being active at the beginning of the simulation.
general	diffusion-grid-size	A positive integer	The size, in number of voxels, of the diffusion grid.
general	diffusion-voxel-size	A positive integer	The size, in µm, of each diffusion voxel.
general	shoving	A Boolean	Whether shoving is active.
general	acquired-mutation-probability	A number between 0 and 1	The probability for a node to mutate every time a new cell is created.
general	close-group-placement	A Boolean	Whether cells from the same subpopulation start the simulation physically close, otherwise they are placed at random.
general	diffusion	A Boolean	Whether diffusion simulation is enabled.
general	step-delay	A positive integer	Milliseconds the simulation will wait in between time steps, can be tuned for performance or monitoring.

Section	Setting	Value	Explanation
general	snapshots	0 or a positive integer	Period of time steps in which snapshots of the graphical representation are taken. If 0, snapshots are deactivated.
general	distributor-2d	One of 'random', 'square' or 'hexagonal'	Which placement strategy will be used for 2D simulations
general	distributor-3d	One of 'random', 'square' or 'hexagonal'	Which placement strategy will be used for 2D simulations
general	cache-positions	A Boolean	Whether initial positions of cells in random placement should be cached for performance.
general	network	One of 'sync', 'async', or 'prob'	The type of network to use, synchronous, asynchronous or probabilistic.
gene-status-data	parse	A Boolean	Whether to parse *Ingres* transcriptomics data
gene-status-data	id	A string	The identifier for the transcriptomics data
gene-status-data	spatial	A Boolean	Whether the transcriptomics data is spatial.
gene-status-data	single-cell	A Boolean	Whether the transcriptomics data should be imported by single cell. Otherwise it is imported by cluster.
mutations	group	Group of subsettings	Each group contains the information for one subpopulation.
group	name	A string	The name of the subpopulation.

Section	Setting	Value	Explanation
group	percent	0 or a positive integer	The percentage of the total number of cells this subpopulation constitutes.
group	mutation	Group of subsettings	Each group contains the information for one node mutation
mutation	node	A string	The name of the node to be mutated.
mutation	sign	A Boolean	The mutation will be over-expression if positive, deletion or under-expression if negative.
mutation	p	A number between 0 and 1	The probability for this mutation, only taken into account if network=prob.
diffusibles	diffusible	Group of subsettings	Each group contains the information for one diffusible substance.
diffusible	name	A string	The name of the diffusible substance.
diffusible	type	One of 'nutrient', 'factor' or 'drug'	The type of substance.
diffusible	input-node	A string	The name of the node that becomes activated by this substance.
diffusible	initial-concentration	A positive number	The concentration of the substance in mM at the beginning of the simulation.
diffusible	diffusion-coefficient	A possitive number	The diffusion coefficient, in m^2/s, of the substance.

Section	Setting	Value	Explanation
diffusible	threshold	A positive number	The concentration in mM above which the substance activates the corresponding nodes of cells in the same voxel.
diffusible	consumption	A positive number	The consumption of each cell, in mol/s/cell.
diffusible	production	A positive number	The production of each cell, in mol/s/cell, only taken into account if type=factor.
stats	filename	A string	The name for the output data file.
stats	stat	A string	The name of the feature to be measured.
timepoint-stats	stat	Group of subsettings	Each group contains the information for one time point measure
stat	type	A string	The name of the feature to be measured every time point.
stat	timepoint	A integer or a range of integers	The timepoints in which measures are to be taken.

Table 4.1: A full description of every setting in *microCJ*

5

Analysis of EGFR heterogeneity using *microCJ* with experimental and transcriptomics data

Contents

5.1 *In vitro* response 196
5.2 An extended RTK module in the MAPK network . . 203
5.3 *microCJ* simulations with new network 209
5.4 Fitting *microCJ* with single-cell RNA-seq data 215
5.5 Conclusions 233
5.6 Methods 240
5.6.1 *In vitro* methods 240
5.6.1.1 Western blot 240
5.6.1.2 Flow cytometry 241
5.6.1.3 Proliferation assays 241
5.6.2 Importing of transcriptomics data 241

In this chapter, I bring together all the methods and data described until now. First I discuss the experimental work I carried out, which despite not being complete, shows what the next steps of this project would be, since no model is complete without experimental validation. Then I take the network described in chapter 3 and extend it to analyse EGFR and the MAPK pathway in the context of TNBC.

I analysed this network using the same attractor methods described in chapter 3, but also running it in a PBN approach. With the microCJ model described in chapter 4, I test the new network, looking at how different perturbations affect growth and providing a proof of concept for an *in silico* drug panel. Finally, I fit the new network with the transcriptomics data studied in chapter 2 and show the result of several simulations, including an *in silico* drug screen.

5.1 *In vitro* response

During my first year, I performed some preliminary *in vitro* experiments. to reproduce results from the literature and to study EGFR inhibition by the tyrosine kinase inhibitor (TKI) gefitinib. EGFR TKIs competitively block the binding of adenosine 5' triphosphate (ATP) to the intracellular catalytic domain of EGFR[342]. I was, sadly, unable to continue these lines of research due to the challenging conditions during the COVID-19 pandemic. Nevertheless, in this section I present the results I managed to obtain to help painting a fuller picture of EGFR targeting in TNBC.

Cell line	Basal subtype	Histology	Mutations in EGFR pathways	Other mutations
MDA-MB-231	Basal-like B	IDC	*BRAF, KRAS*	*CDKN2A, PDGFR, TP53*
MDA-MB-468	Basal-like A	DC	*PTEN, EGFR amplification*	*RB1, SMAD4, TP53*

e

Table 5.1: Characteristics of triple-negative breast cancer cell lines used in the preliminary experiments. Adapted from Guerrab *et al.* [342]. DC: ductal carcinoma; IDC: invasive ductal carcinoma.

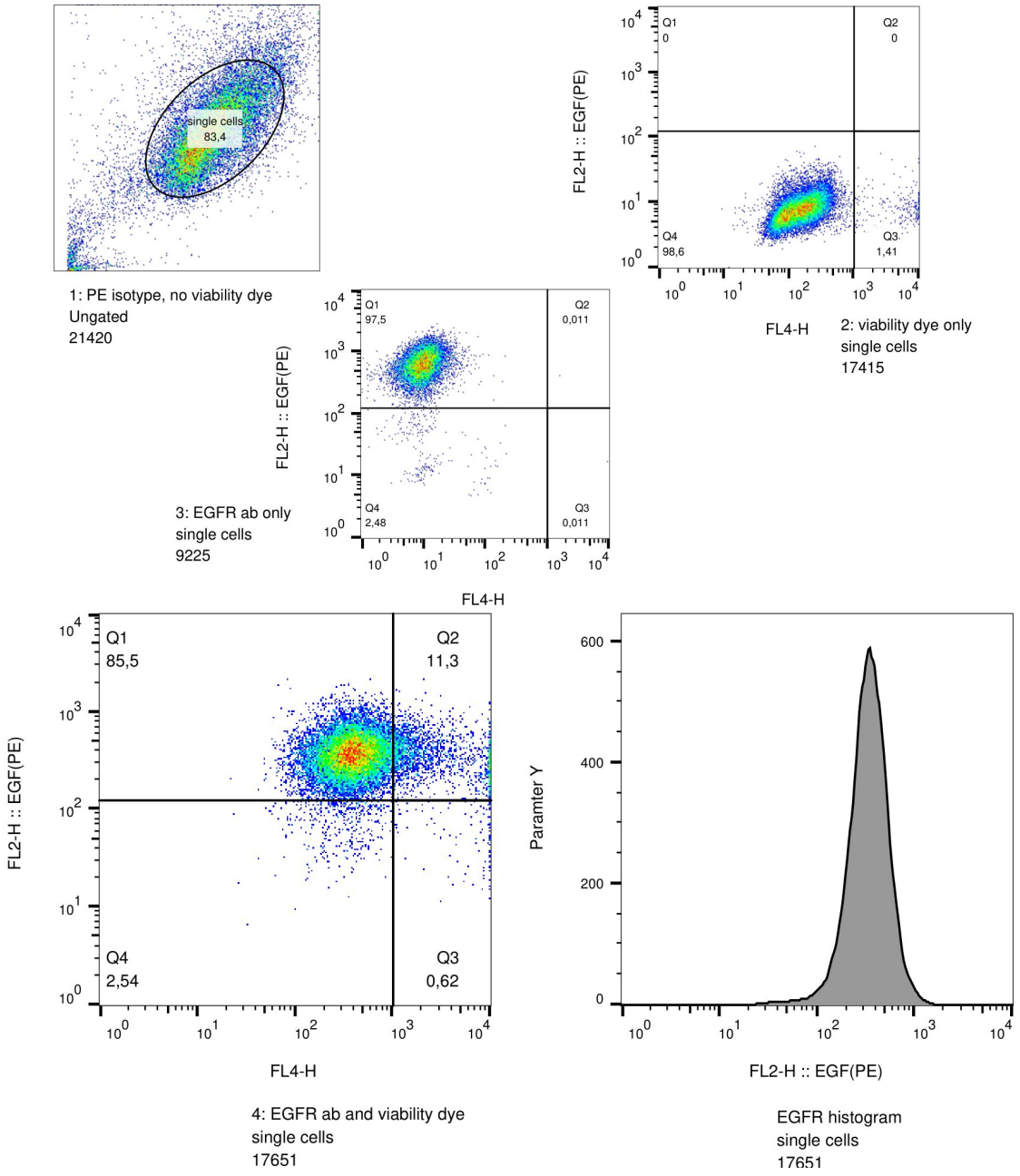


Figure 5.1: Flow-cytometry of MDA-MB-468 cells labelled with an anti-EGFR primary antibody. EGFR antibody from BD Biosciences (catalog no. 555997). The viability assay LIVE/DEAD stain from ThermoFisher (catalog no. L34973) was also used. A million cells were incubated with the antibody and viability assay according to the manufacturers' protocol. FACS was performed right after the incubation in a FACSCalibur machine.

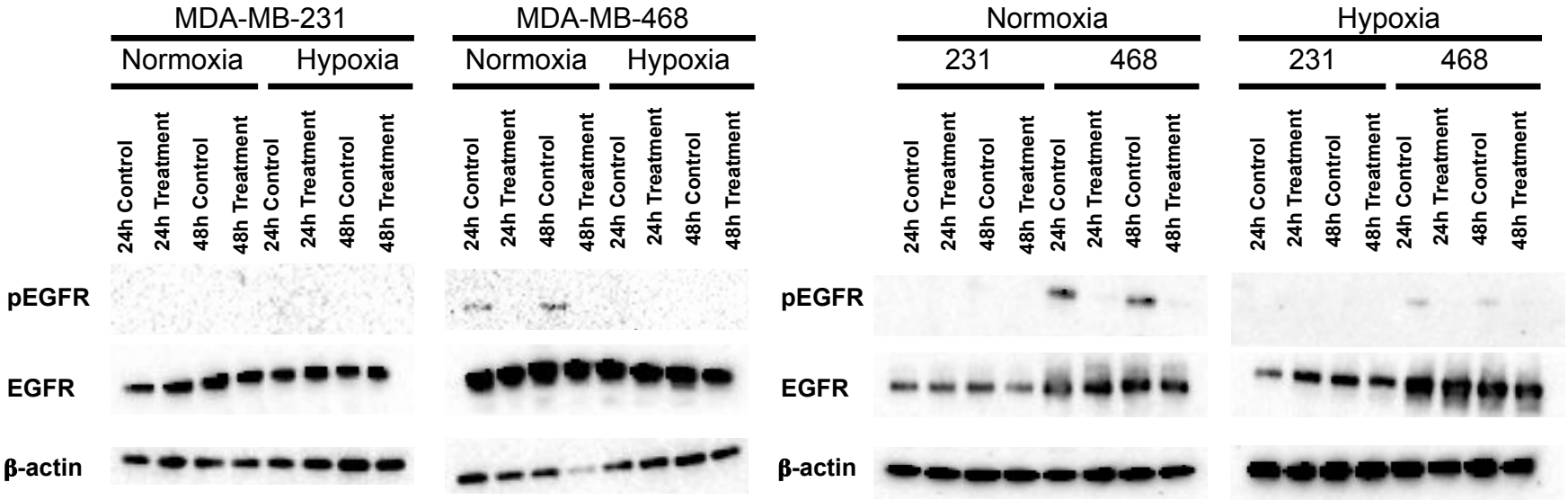
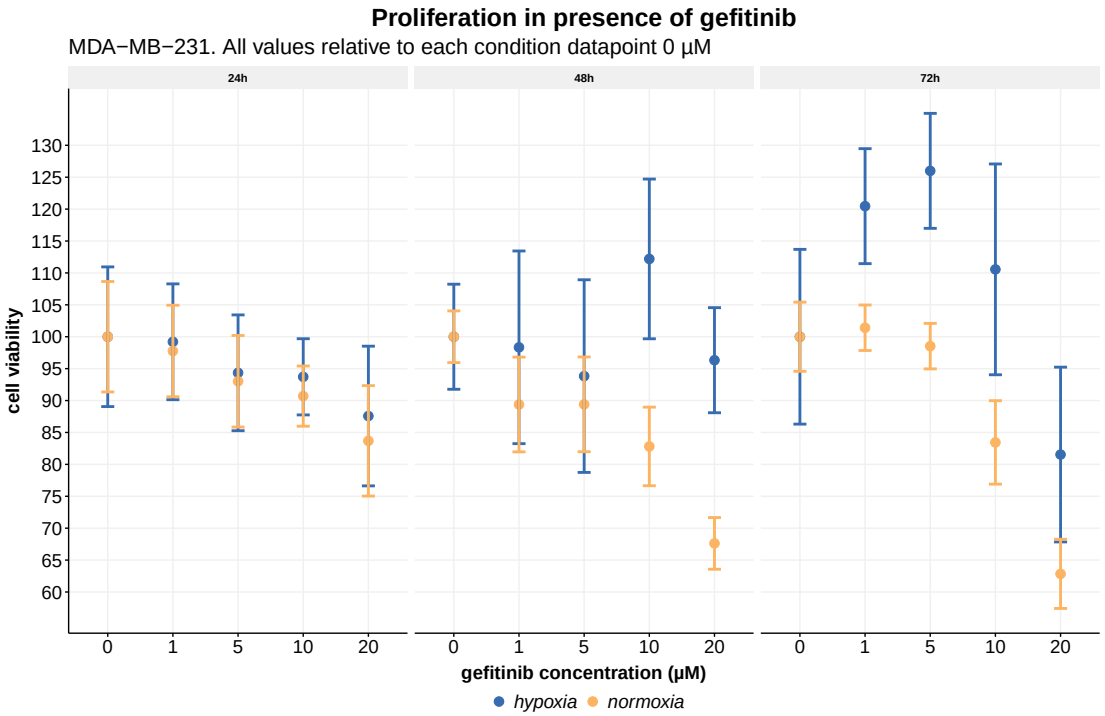
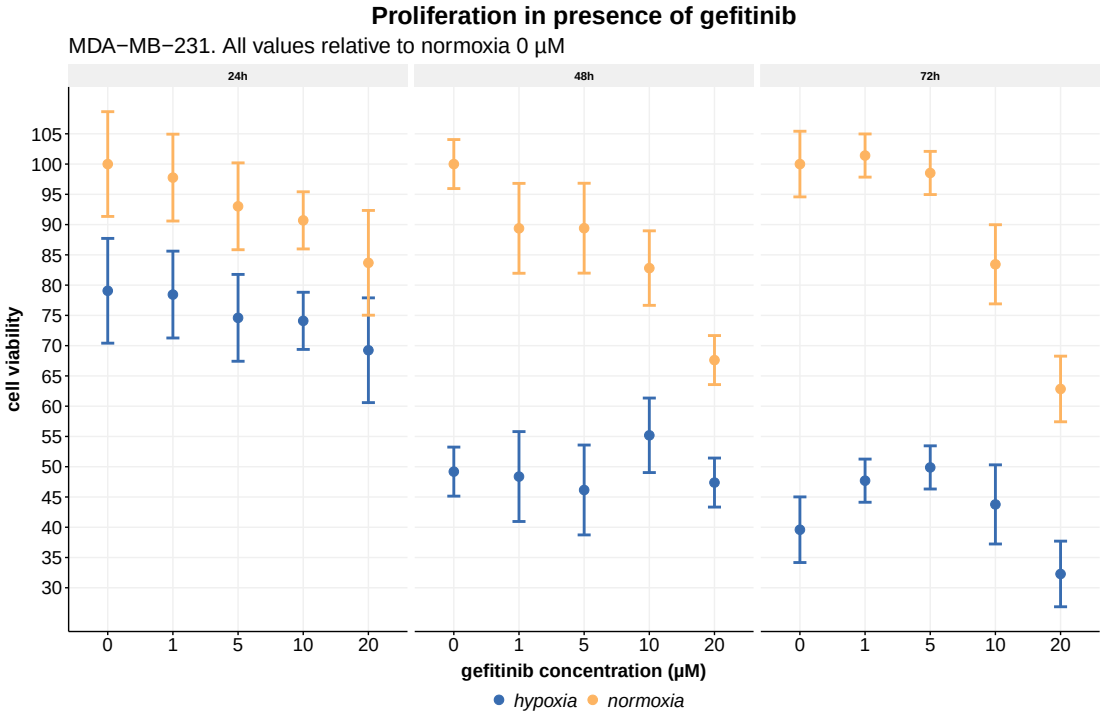
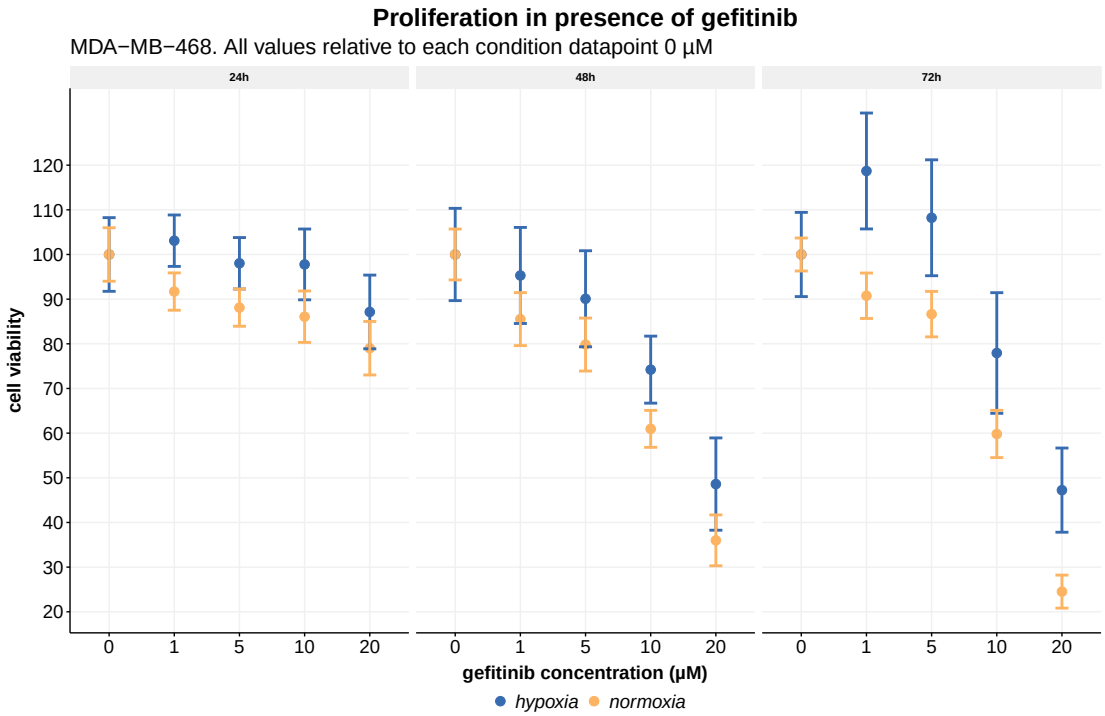


Figure 5.2: Expression of EGFR and p-EGFR in MDA-MB-231 and 468 cells treated with gefitinib. 24 hours after being seeded cells from both lines were treated with gefitinib 10 μM for 24 and 48 hours, both under normoxic and hypoxic conditions. For Western Blot, cells were lysed and the whole cell protein extract was separated and immunoblotted using the indicated antibodies. β-actin antibody was used as a loading control. On the left, samples were divided among two gels by ther cell line, and on the right by the oxygen condition.


Proliferation in presence of gefitinib
MDA−MB−231. All values relative to each condition datapoint 0 µM
24h
48h
72h
cell viability
130
125
120
115
110
105
100
95
90
85
80
75
70
65
60
0
1
5
10
20
gefitinib concentration (µM)
hypoxia
normoxia


Proliferation in presence of gefitinib
MDA−MB−231. All values relative to normoxia 0 µM
24h
48h
72h
cell viability
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
0
1
5
10
20
gefitinib concentration (µM)
hypoxia
normoxia


Proliferation in presence of gefitinib
MDA−MB−468. All values relative to each condition datapoint 0 µM
24h
48h
72h
cell viability
120
110
100
90
80
70
60
50
40
30
20
0
1
5
10
20
gefitinib concentration (µM)
hypoxia
normoxia

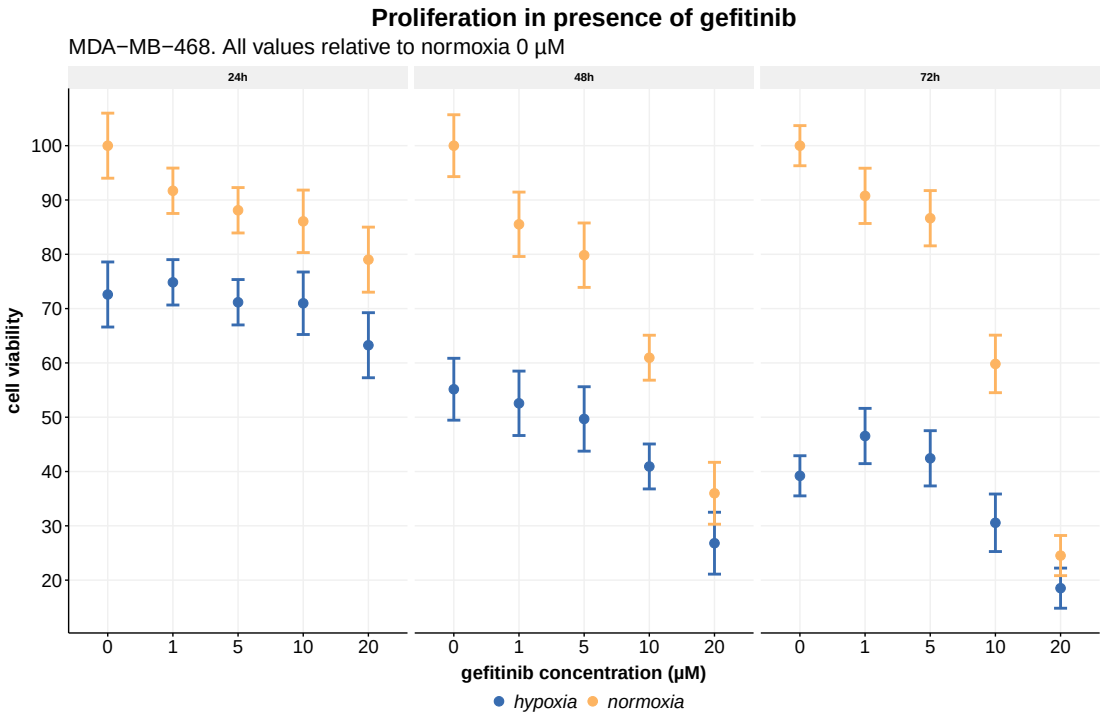


Figure 5.3: Proliferation of MDA-MB-468 and MDA-MB-231 cells treated with gefitinib. Cells were treated with different concentrations of gefitinib, for 24, 48 or 72 hours, and grown under normoxic or hypoxic conditions. Cell number was measured with CyQUANT proliferation assay. Values are either expressed in proportion to normoxia or hypoxia 24h 0μM, respectively (first and third subfigures), or both normoxia and hypoxia values are expressed in proportion to normoxia 24h 0μM (second and fourth. subfigures). Error bars correspond to the standard error of the mean of the three technical replicates.

I used the triple negative breast cancer cell lines MDA-MB-231 and MDA-MB-468, both of them have mutations in pathways where EGFR is involved (Table 5.1). MDA-MB-468 overexpresses EGFR while MDA-MB-231 does not. Therefore, they can function as a basic model for response to EGFR targeting therapy. Additionally, I replicated most experiments in both normoxic and hypoxic conditions to examine the differences caused by hypoxia in cell growth and EGFR activation.

First, I confirmed levels of expression of EGFR in the cell lines, along with the capability of gefitinib to inhibit EGFR. To check EGFR expression, I performed flow cytometry of MDA-MB-468 cells (Figure 5.1) and Western Blot of both cell lines. In the flow cytometry experiment, gating was performed based on PE-isotype negative control, but due to a software error it was not saved. It shows expression of EGFR in living cells (86%, top left quadrant). As expected from the literature[342] MDA-MB-468 cells have a high level of expression of EGFR. Comparative analysis of EGFR expression in both cell lines was planned, in order to assess the relative expression of the protein, using a cell line with no EGFR expression as a biological negative control, but could not be completed.

Western Blot analysis (Figure 5.2) was intended to confirm inhibition of EGFR by gefitinib, and to compare its effect under normoxia and hypoxia. Cells were treated with gefitinib 10 μM for 24 or 48 hours. Specifically, I was trying to reproduce the results from Guerrab *et al.*[342].

Immunoblotting for p-EGFR proved challenging, and the only consistent result was inhibition of EGFR by gefitinib in MDA-MB-468, both for 24 and 48 hours treatments. One of the samples showed inhibition in hypoxia but only normoxia gave the same result for every experiment. The p-EGFR antibody is not very sensitive, and because MDA-MB-231 does not overexpress EGFR (as it is hinted by in the anti-EGFR antibody blot), there was not any visible staining of these cells, with or without treatment. Still, for MDA-MB-468, it can clearly be seen how gefitinib inhibits the phosphorylation of EGFR, both at 24 and 48 hours, and both in normoxia and hypoxia, although the signal in normoxia is more intense.

Then, I studied the effect of gefitinib on cell proliferation in the two cell lines (Figure 5.3). In order to do this I cultured cells under normoxia or hypoxia and treated them with several gefitinib concentrations. It can be seen that MDA-MB-468 cells are more sensitive to EGFR-targeted treatment, and both cell lines are affected by hypoxia.

The results of these experiments show a clear effect of gefitinib on proliferation, such as reported by Guerrab *et al.* [342]; and, perhaps more interestingly, a differential response to gefitinib on cells grown under hypoxia. Hypoxic cells seem to be less sensitive to gefitinib, although this need to be further studied.

5.2 An extended RTK module in the MAPK network

In order to better capture the EGFR downstream and the possible mechanisms of resistance, I extended the MAPK network I use in *microCJ* (figure 5.4). I added the receptors AXL, MET and IGF1R, which in the context of the MAPK and AKT/PI3K/mTOR pathways activate similar targets[341], which is abstracted as the node "RTKs crosstalk", that activates what were EGFR targets in the previous network. This node will be active when any of the RTK is active; the Boolean rule for it is: *EGFR or AXL or MET or IGF1R*. EGFR has now two inputs, ATP binding and EGF binding, representing those two binding sites. ATP binding is active unless an EGFR TKI is introduced in the simulation, EGF binding will be active if there is EGF in the vicinity of the cell, and it can be blocked by an anti-EGFR monoclonal antibody (mAb). For the other receptors, input nodes have been added (IGF, GAS6 and HGF) as well as targeted drugs (IGF1Ri, AXLi and METi). Additionally, a PI3K inhibitor drug node was also added. Figures 5.5 to 5.9 show the network analyses analogous to the ones presented in chapter 3: attractor search for synchronous updates of the network.

The wild type (no perturbations other than inputs set to off) only presents proliferation attractors, but no apoptosis (figure 5.5, which demonstrates a higher tendency towards proliferation than the original network, which did have attractors

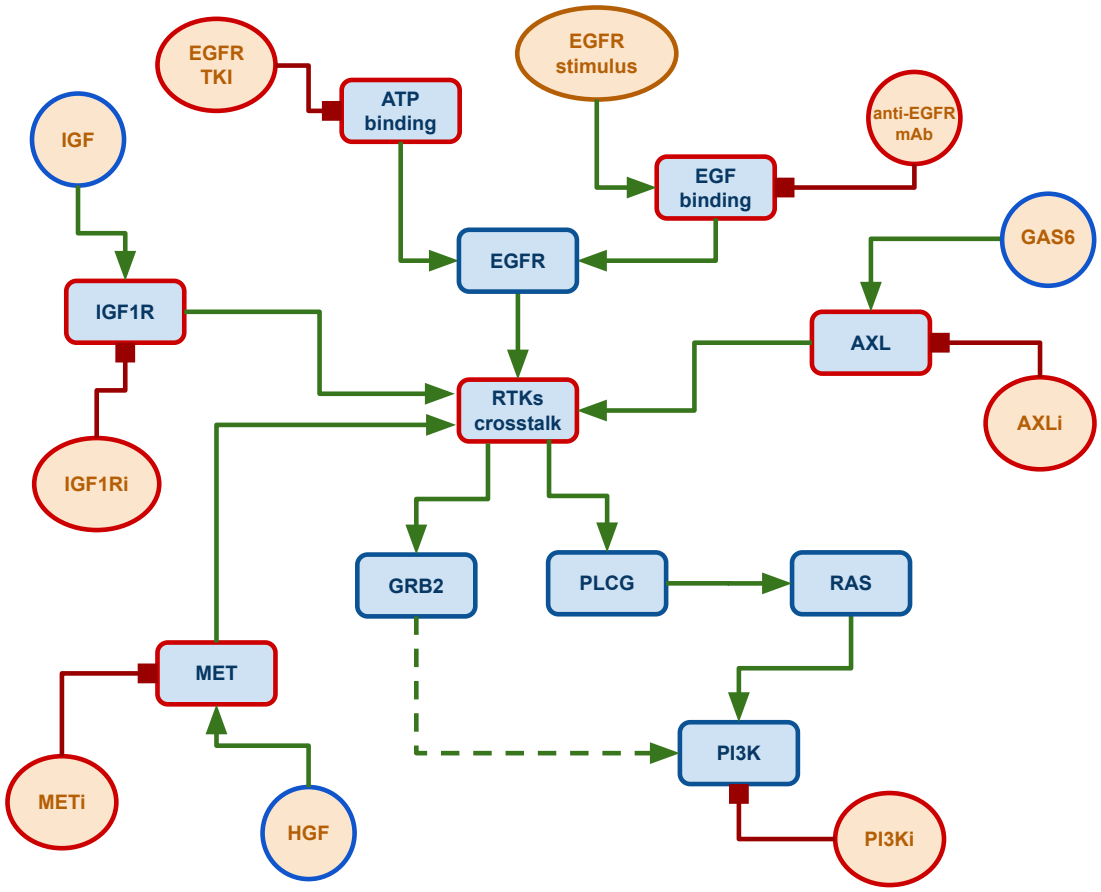


Figure 5.4: New nodes added to the MAPK network. Nodes circled in red are the new nodes added to the network. Nodes with blue background represent genes or proteins, nodes with orange background represent inputs. Green arrows represent positive or activating interactions, red square arrows represent negative or inhibiting interaction. The dashed arrow indicates that there are other nodes in between that are not represented in this diagram.

for apoptosis. This is likely caused by the fact that there are now four different nodes that activate the proliferation pathway, instead of just EGFR. If any of them is active in the random initial state, it will probably reach a proliferation attractor. The EGFR over-expressed network (figure 5.6) presents attractors with all of these three fates, as the original EGFR+ network did (figure 3.7). The p53 attractors (figure 5.7) are also consistent with those for the original network.

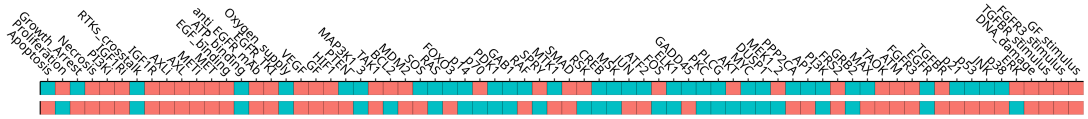
Figure 5.8 shows the steady states for the inhibitor drug nodes EGFR TKI and a combination of EGFR TKI and METi. Both of them prevent the apparition of proliferation attractors, but without apoptosis or growth arrest.

Another comparison with the original network analysis is presented in figure 5.9,



Figure 5.5: 8-state synchronous attractors for the unperturbed extended network. Every row of tiles corresponds to one state, in green for nodes that are active (one) and red for nodes that are inactive (zero). Every block of states represent an attractor. Each of them is made up of eight states that transition to the next one in a loop.

which shows the same perturbations as in figures 3.10 and 3.11, results are similar in both cases.



(a)

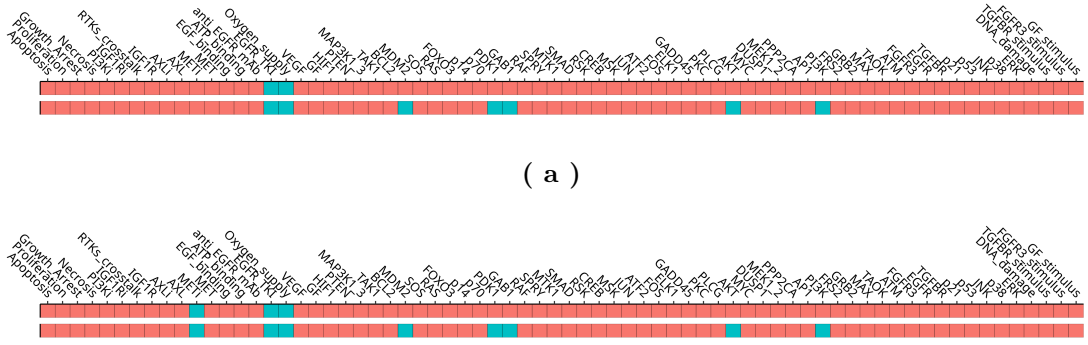


(b)

Figure 5.6: Attractors for the EGFR over-expressed extended network. Stable states ((a)) and 8-state attractors for the network with EGFR fixed to one (always active). See figure 5.5 for an explanation of the figure.



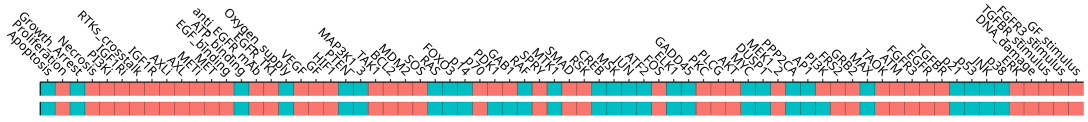
Figure 5.7: 8-state attractors for the p53- extended network. Stable states (The network had p53 fixed to zero (never active). See figure 5.5 for an explanation of the figure.



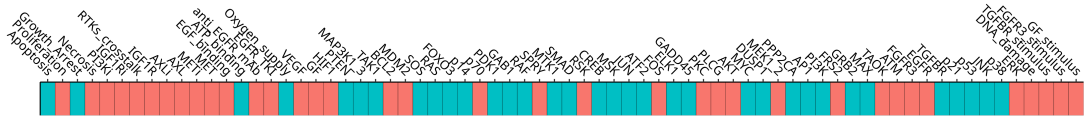
(a)

(b)

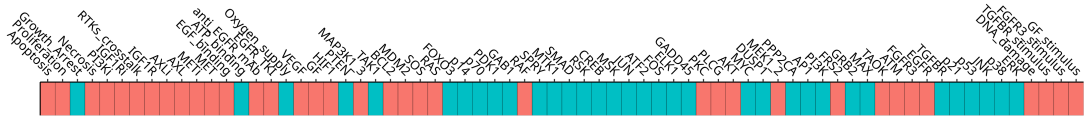
Figure 5.8: Attractors for positive drug inputs in the extended network. (a): EGFR TKI, (b): EGFR TKI and METi. See figure 5.5 for an explanation of the figure.



(a)



(b)



(c)

Figure 5.9: Attractors for different perturbations of the extended network. (a): RAS+, JNK+. (b): TGFBR+. (c): ERK+, TGFBR+. See figure 5.5 for an explanation of the figure.

5.3 *microCJ* simulations with new network

In this section I show a few results from *microCJ* simulations using the new networks. In these simulations, the network was run as a probabilistic one, where I introduced a difference to the synchronous network analysed in the previous section. In the probabilistic extended network, the RTK crosstalk node is not governed by an OR rule with all the RTKs, but instead it has a different rule for each of them (active if EGFR=1, active if MET=1, and so on). The EGFR rule has a p=0.4, while the others have a p=0.2. This is to represent the comparative importance that EGFR has in our model, in the data used and in the assumptions of the original network. Using probabilistic rules instead of an OR one intends to reign the bias towards proliferation in, by not activating the node RTK crosstalk anytime that any of the RTKs is active. This actually models the fact that crosstalk between these RTKs is necessary, activation of the downstream pathway is not performed independently. Doing it probabilistically allows more flexibility in our assumptions than just adding AND rules.

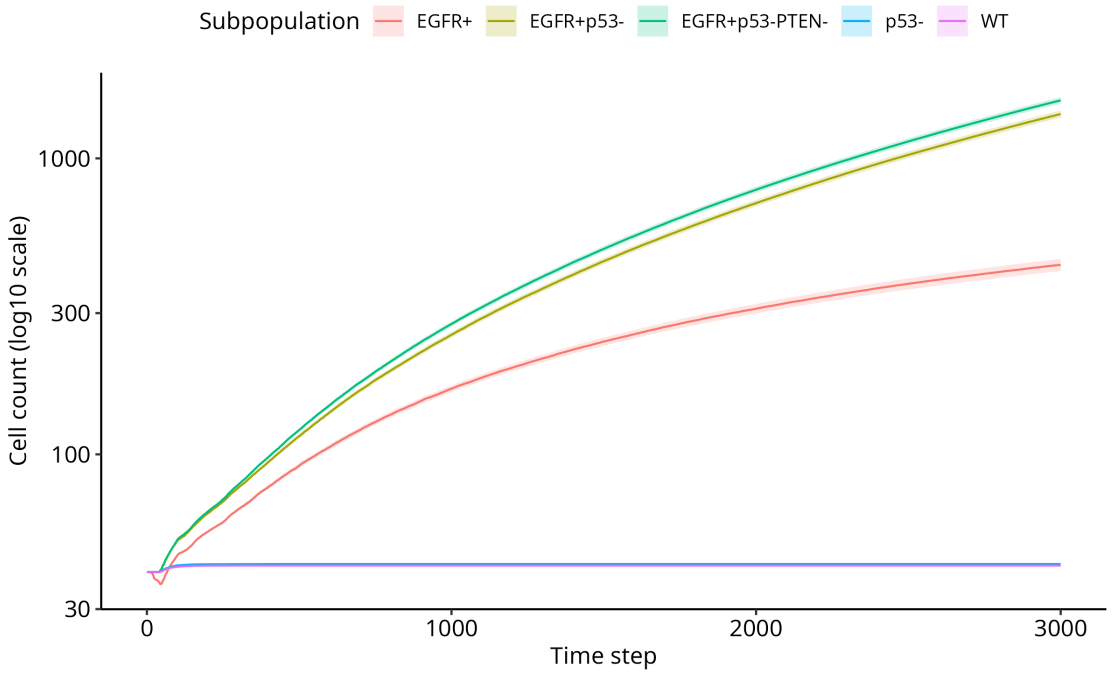
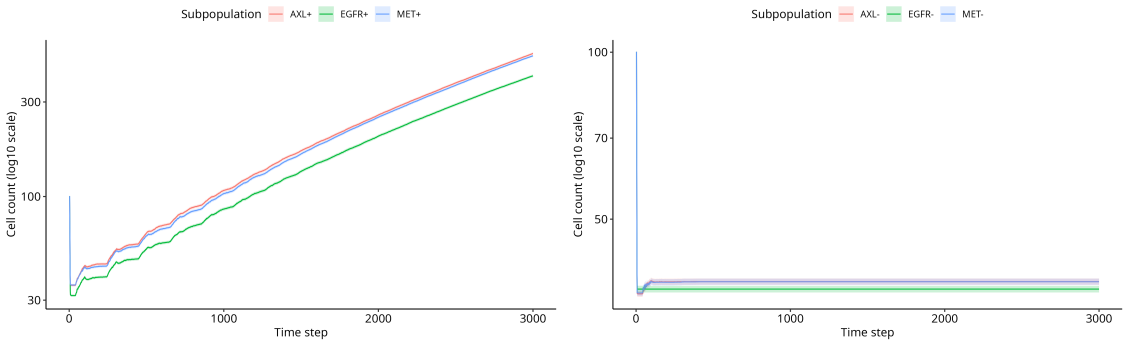


Figure 5.10: Competition simulation with the new network. Simulations were run for 3000 steps with 100 replicates, in a 2D space. The figure shows the average amount of cells at each time step per subpopulation and the standard error of the mean.



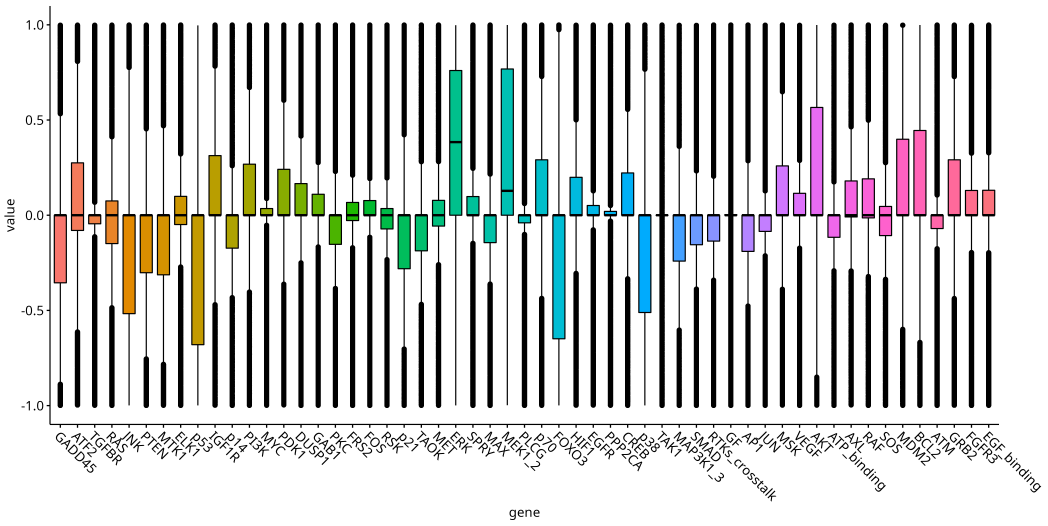
(a) (b)

Figure 5.11: *microCJ* running the new extended network with different perturbations. Populations were simulated independently, with 100 replicates each. See figure 5.10 for more details.

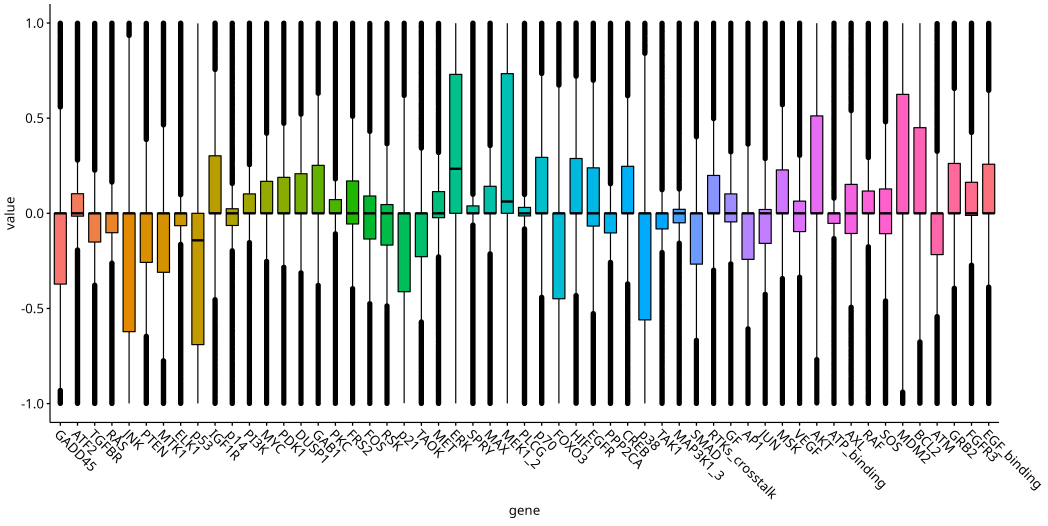
Figure 5.10 displays the result of a *microCJ* competition simulation with the usual subpopulations, using the extended network. In this simulation, the EGFR+ subpopulation is more proliferative compared to the simulation with the original network. The total numbers reached by the subpopulations are lower than in the simulations with the original network because the simulations in this section were run in 2D space instead of 3D. This was done to reduce the computational load, which increases with every node added to the network inside each cell. Nevertheless, the relative proliferation rate of each subpopulation compared to the others can be compared across the different simulations.

The RTKs AXL, EGFR and MET were overexpressed or deleted in figure 5.11 a and figure 5.11 b , respectively. Interestingly, the EGFR+ population has a lower proliferative rate than the others, this is probably due to the activation of nodes that block EGFR (refer to figure 3.3 for the complete view of the network) whereas there are no nodes blocking the other RTKs, which I thought would be countered by their lower probability in the RTK crosstalk node rules. On the other hand, in the deletion simulations all populations have almost identical growth curves.

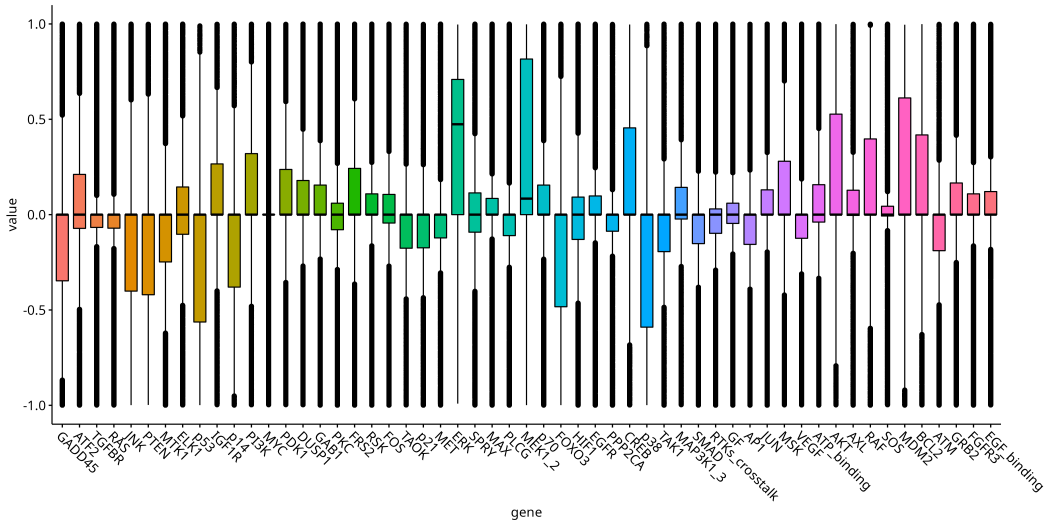
I then performed the acquired mutations experiments, as in figure 4.19, with the extended network. Four types of simulation were run, one with WT cells, another with EGFR TKI, a third one with EGFR mAb and the last one with EGFR TKI + METi (figure 5.12). These inhibitor nodes are activated by the presence of a drug,


value
1.0
0.5
0.0
-0.5
-1.0
GADD45
ATF2
TGFBR
RAS
JNK
PTEN
MTK1
ELK1
P53
IGFR
P14
PI3K
MYC
PDK1
DUSP1
GAB1
PKC
FRS2
FOS
RSK
P27
TAOK
MET
ERK
SPRY
MAX
MEK1_2
PLCG
P70
FOXO3
HER1
EGFR
PP2CA
CREB
P38
TAK1
MAP3K1_3
SMAD
RTKs_crosstalk
GF
AP1
JUN
MSK
VEGF
AKT
ATF2_binding
AXL
RAF
SOS
MDM2
BCL2
ATM
GRB2
EGFR3
EGF_binding
gene

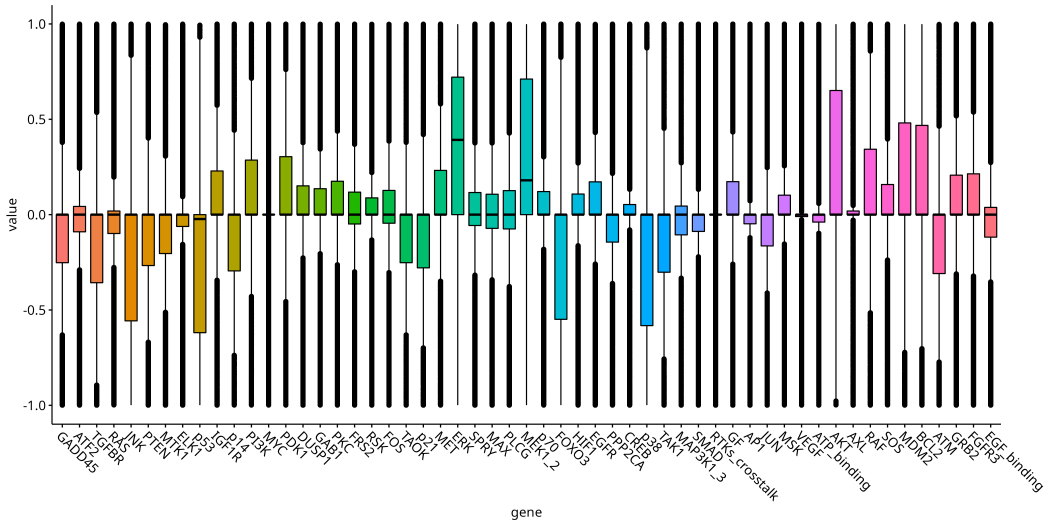
(a)


value
1.0
0.5
0.0
-0.5
-1.0
GADD45
ATF2
TGFBR
RAS
JNK
PTEN
MTK1
ELK1
P53
IGFR
P14
PI3K
MYC
PDK1
DUSP1
GAB1
PKC
FRS2
FOS
RSK
P27
TAOK
MET
ERK
SPRY
MAX
MEK1_2
PLCG
P70
FOXO3
HER1
EGFR
PP2CA
CREB
P38
TAK1
MAP3K1_3
SMAD
RTKs_crosstalk
GF
AP1
JUN
MSK
VEGF
AKT
ATF2_binding
AXL
RAF
SOS
MDM2
BCL2
ATM
GRB2
EGFR3
EGF_binding
gene

(b)



(c)



(d)

Figure 5.12: Acquired mutations in the extended network under different treatments. Simulations were run with 500 initial cells, all of them WT, and the acquired-mutation-probability parameter set to 0.8, using the extended network (figure 5.4). (a) shows the control simulation, (b) a simulation under treatment with an EGFR TKI, (c) with an EGFR mAb, and (d) with a combination of EGFR TKI and METi. Simulations were run for 10000 steps, and the drug treatment lasted from time step 200 to time step 6000. The figure summarises the data points from all cells in the 100 replicates. In the box plots, the mid line is the median value, the lower and upper hinges correspond to the first and third quartiles, and the whiskers extend to the largest and smallest value, respectively, no further than 1.5 * inter-quartile range. Data beyond the end of the whiskers are plotted individually.

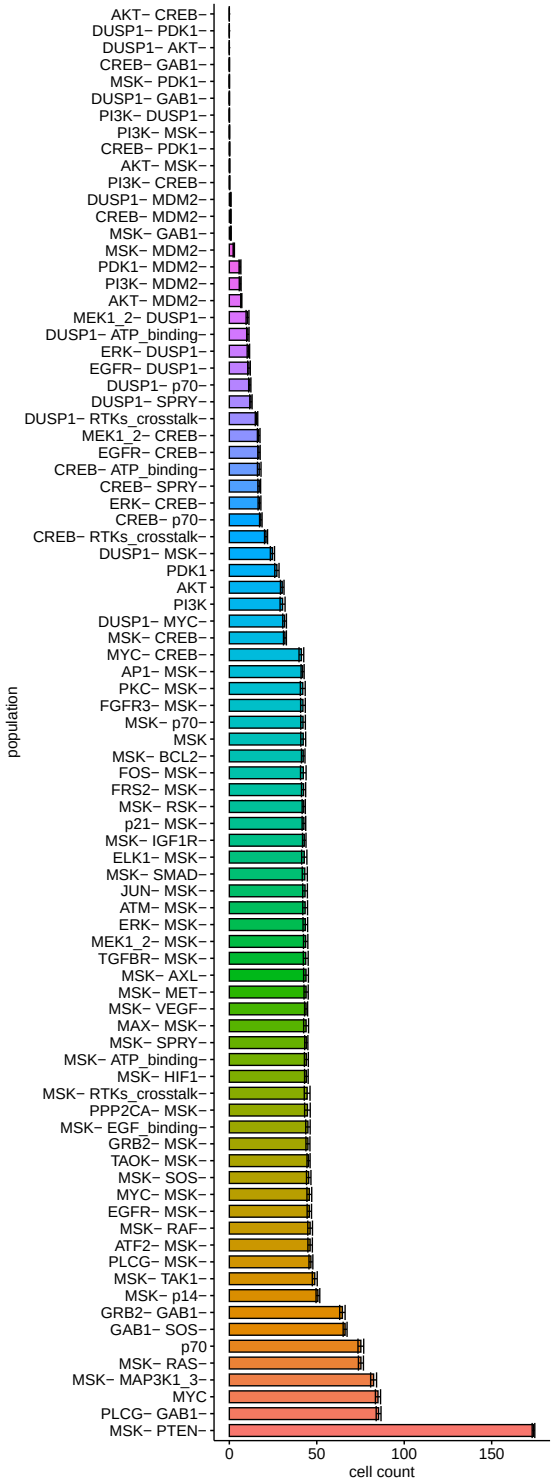


Figure 5.13: *microCJ* mutation panel simulations. Each population with a set of inactivating mutations was simulated independently for 3000, with 100 replicates. The figure shows the average amount of cells at the end of the simulation, the error bars represent the standard error of the mean.

as described in chapter 4, that was injected into the simulation from time step 200 to 6000, for a total simulation duration of 10,000 steps. There are not many differences among them nor with the original experiment. As expected, p53 and PTEN are consistently downregulated, and ERK, MEK and AKT, among others, are upregulated. The new RTKs nodes, IGF1R, MET and AXL display a very modest upregulation, or none at all in the case of MET. The trend continues to be that the more downstream nodes — closer to the proliferation node — are more upregulated than the input or directly downstream of input nodes. In the simulations with RTKs inhibitor nodes active, the overexpression of the RTKs nodes is a bit higher, which makes sense since there is this inhibiting pressure against them, but not very much.

Finally, I performed several experiments that measured the effect of these TKIs drugs, anti-EGFR mAb, and PI3K inhibitors on wild type cells proliferative rate, but none of them had a significative effect (data not shown). This may be due to the bigger model response to donwstream targeting, as discussed above. Drug modelling in *microCJ* also needs some improvement, as discussed in chapter 4.

Due to this setback, I decided to instead test the effects of deletion for every node in the network and every pair of nodes to elucidate which targets or combination of targets were the most promising. First, I switched off every node and pair of nodes in the network, and for each of these perturbations I computed the attractors of the perturbed network. I kept all those perturbations that produced an apoptosis or growth arrest attractor. If a single node perturbation had the same attractor that a pair of nodes, only the single node was kept. If a single node perturbation produced a growth arrest attractor but a combination that contained that node produced an apoptosis attractor, only the pair was kept. In other words, for a list of perturbations and fate present in the attractors such as:

A - Growth arrest

A & B - Apoptosis

C - Apoptosis

C & A - Apoptosis

only the perturbations A&B and C would be kept. Then, I ran *microCJ* simulations

for the selected perturbations. The results of this are shown in figure 5.13. A separate simulation was run for every node or pair of nodes, with WT cells, 3000 steps and 100 replicates. Finally, I chose the perturbations that resulted in a average cell count at the end of the simulation of 35 or less — that is, everything with a bigger response than that of MSK-CREB-, which left a selection of 36 perturbations out of 85. This selection is the basis for the *in silico* drug panel in the next section. To create the drugs targeting the selected nodes, some more processing needed to be done. Wherever the node RTK crosstalk was present in a perturbation, it was substituted by a combination of EGFR TKI and METi, which has been used in several TNBC clinical studies with promising results[137,138,343,344]. The EGF binding node was substituted by anti-EGFR mAb and ATP binding by EGFR TKI. For each node present in this final selection, a drug input node inhibiting that node was created, allowing the simulation of these drug combination treatments within the drug diffusion framework.

5.4 Fitting *microCJ* with single-cell RNA-seq data

The single-cell and spatial RNA-seq datasets described and analised in chapter 2 were use to fit the extended network described previously, producing a PBN for each cell in each of those datasets, using *Ingres* (chapter 3). The single-cell PBNs were imported (as described in chapter 4) and used for several simulations in *microCJ* (figure 5.15 a). Figure 5.15 b displays the growth curves of the subpopulations in the control dataset, with the data as processed by the authors of Savage *et al.*[1]. Here it can be seen that the cluster labelled as proliferation is indeed the one with the highest, by a large margin, proliferative rate in the model. After proliferation, the mesenchymal/stem cluster is the biggest one at the end of the simulation, but it does not have a high proliferative rate, it just goes into apoptosis less frequently, which is also what one could expect of it. This shows that, at least in this very general way, certain phenotypes of the cell clusters can be preserved through *Ingres* and recapitulated in *microCJ*: only proliferative vs cell death phenotypes are present in the current model.

Table 5.2: List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Savage *et al.*[1] dataset.

Drug target	**Subpopulations**
AKT	C-B2, C-Prolif1
AKT CREB	C-B1, C-B2, C-Prolif2
AKT MDM2	C-B2, C-Prolif2
AKT MSK	C-B1, C-B2, C-MSL-EGFRhi, C-Prolif2, T-LAR
CREB GAB1	C-Prolif1
CREB MDM2	C-B3
CREB p70	C-Prolif1
DUSP1 AKT	C-B2, C-MSL-EGFRhi, C-Prolif2, OT-Prolif
DUSP1 EGFR MET	C-Prolif1
DUSP1 MDM2	OT-Prolif
DUSP1 MYC	C-Prolif1
DUSP1 p70	C-MSL-EGFRhi
ERK CREB	C-B3, OT-Prolif
ERK DUSP1	C-B3, C-Prolif1
MEK1_2 DUSP1	C-Prolif1
MSK CREB	C-Prolif1, T-LAR
MSK GAB1	C-B3
MSK MDM2	C-Prolif2, T-LAR
MSK PDK1	T-LAR
PI3K	C-B1, T-M, T-MSL-EGFRhi
PI3K CREB	OT-Prolif, T-M, T-MSL-EGFRhi
PI3K DUSP1	C-MSL-EGFRhi, OT-Prolif, T-M, T-MSL-EGFRhi
PI3K MDM2	C-B1, C-B3, C-MSL-EGFRhi, T-M, T-MSL-EGFRhi
PI3K MSK	C-B1, T-LAR, T-M, T-MSL-EGFRhi

Table 5.3: List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Gambardella *et al.*[297] dataset.

Drug target	**Subpopulations**
AKT CREB	HCC38, MDAMB468
AKT MDM2	HCC38
AKT MSK	HCC1187, HCC1937, HCC70, MDAMB468
CREB GAB1	BT20
CREB MDM2	HCC38
CREB PDK1	BT20, MDAMB468
DUSP1 AKT	HCC38, MDAMB468
DUSP1 MSK	MDAMB436
DUSP1 PDK1	MDAMB468
ERK CREB	HCC1937, HCC70
ERK DUSP1	HCC1937, HCC70
MSK CREB	HCC38, MDAMB436
MSK GAB1	HCC1187, HCC1937, MDAMB436
MSK MDM2	HCC1187, HCC1937, HCC70, MDAMB436
MSK PDK1	HCC1187
PI3K	BT549, CAL851
PI3K CREB	BT549, CAL851
PI3K DUSP1	BT20, BT549, CAL851
PI3K MDM2	BT20, BT549, CAL851
PI3K MSK	BT20, BT549, CAL851, HCC1187, HCC70, MDAMB436

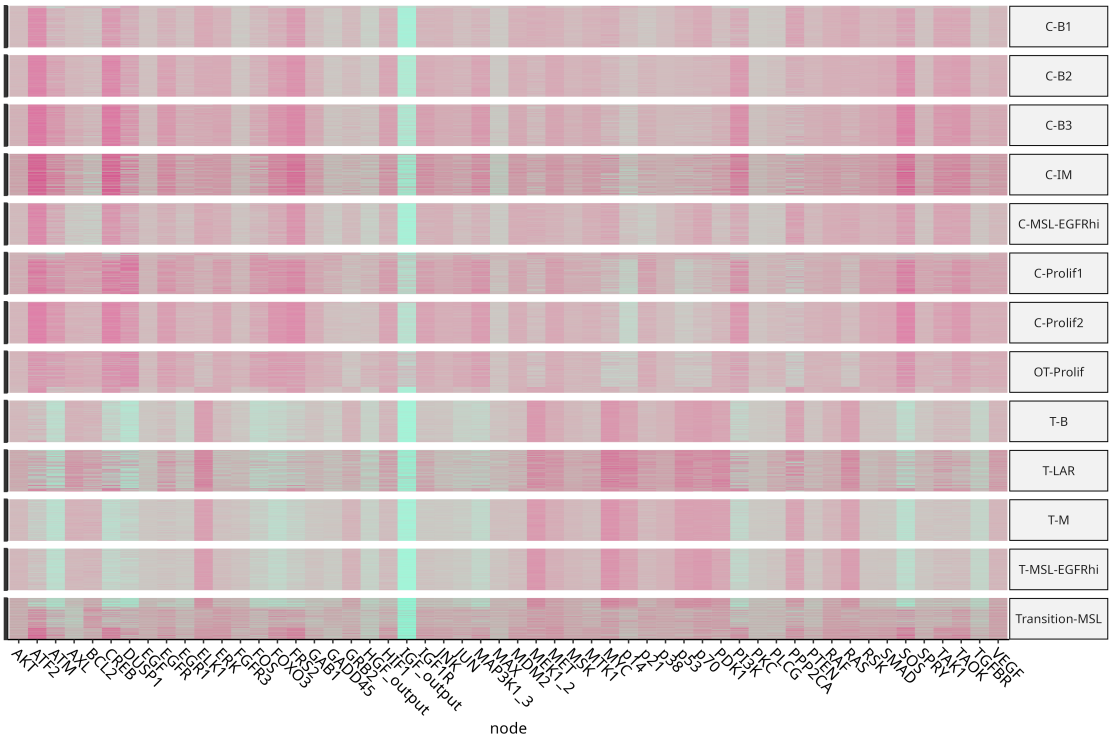
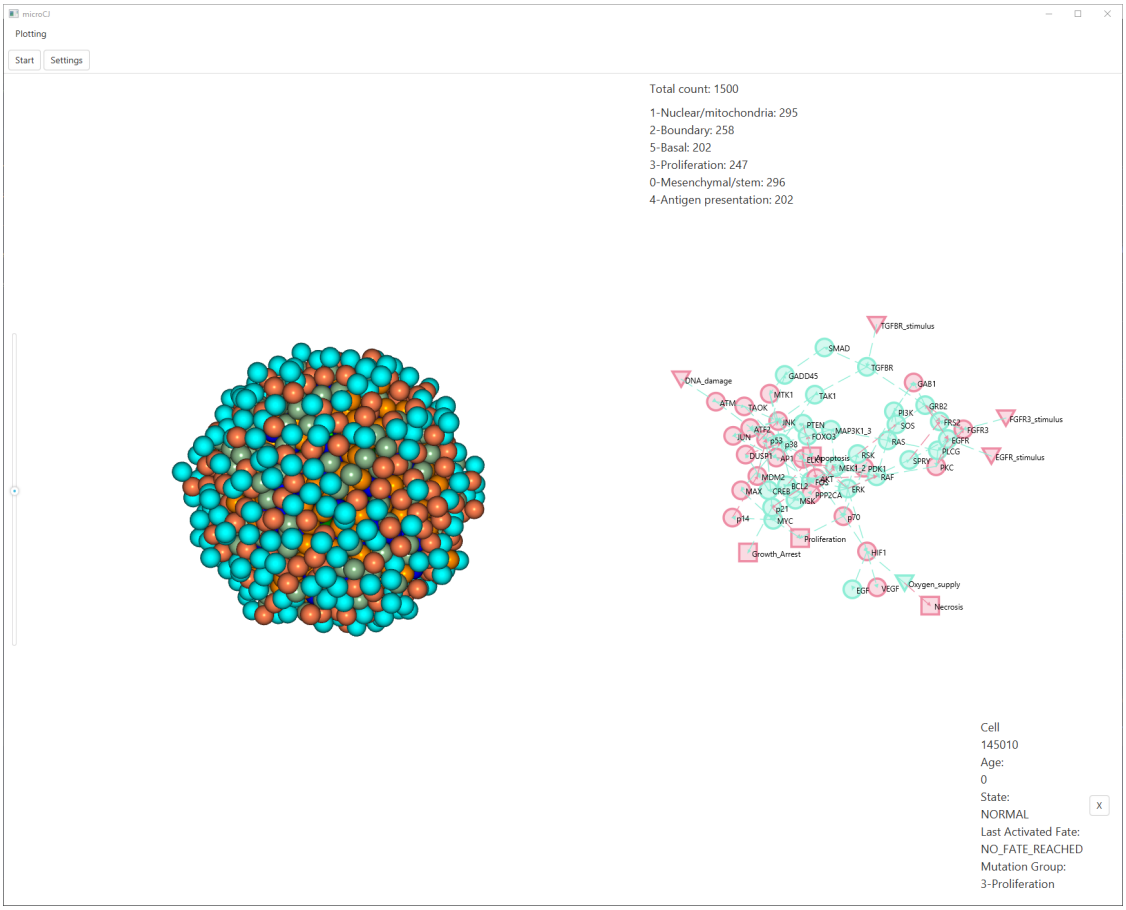


Figure 5.14: An overview of the PBNs produced by *Ingres*, one per each cell, with the dataset from Savage *et al.*[1]. The heatmap shows the probabilities for the fixed function (to one or to zero) for each node, as computed by *Ingres*. Each row correspond to a single cell. Cells are grouped by cluster.

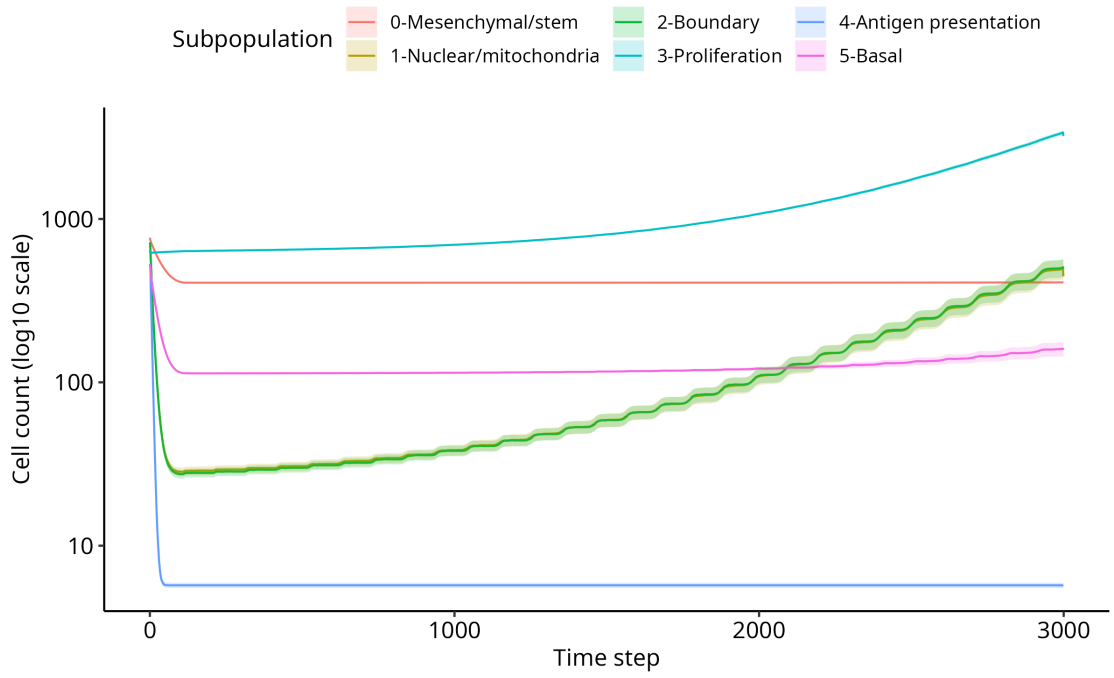
As explained in chapter 2, I performed my own analysis of the Savage *et al.*[1] datasets, merging the control, treatment and off-treatment data. After generating the single-cell PBNs with *Ingres* (figure 5.14) and loading them into *microCJ*, they were run in competition in a simulation of which the results are shown in figure 5.16. The off-treatment (OT) proliferation cluster is the one with the highest proliferative rate, which proves again the ability of this pipeline to roughly preserve the proliferative phenotypes. Most other populations quickly die off, except for C-Prolif2 (as it would be expected), C-B2, T-LAR, and T-MRL-EGFRhi. In general, most cells at the beginning of the simulation will die shortly after the start, and only a few will proliferate, responsible for most of the size of the final tumour. This way, the cells with the most extreme protein normalised enrichment scores (NES) in VIPER, the ones with the higher scores for proliferative genes and lowest scores for tumour suppressor genes will be selected. This is the expected

behaviour of the model in the current way it is set up, but more granularity in their phenotype would be desired.

Then, a drug panel was simulated in *microCJ*. For every drug or drug combination in the list, selected as explained in the previous section, an *microCJ* simulation was run. The drug treatments lasted for the entirety of the simulation. In each simulation, all clusters in the scRNA-seq dataset were run. Figure 5.17 shows the results of this panel, relative to the control. One can see that not every subpopulation is sensitive to the drugs involved — for example, for the most part, the OT-proliferation cluster does not display sensitivity to any of the combinations. Table 5.2 is a list of the predicted top five most effective drugs for each subpopulation. Figure 5.18 plots the expression of these targets in Seurat. One can see that these combinations can be classified in five groups: combinations containing one of AKT, CREB, DUSP1, MSK or PI3K. Many of these targets have been already tested in



(a)



(b)

Figure 5.15: Fitting the model with the control scRNA-seq dataset from Savage *et al.*[1]. All populations, imported by single cell from the control Savage *et al.*[1] dataset, processed by the authors of the article, were simulated in competition for 3000 steps with 100 replicates. (a) shows the*microCJ* GUI with the scRNA-seq data loaded. (b) shows the average amount of cells per population at each time step, and the standard error of the mean.

the clinical setting, especially PI3K and AKT. What is novel about this method is the specific combinations proposed by the model. In the discussion below I will go over each combination and what the next steps might look like.

The same pipeline of simulations was performed with the cell line scRNA-seq from Gambardella *et al.*[297]. The PBNs produced with *Ingres* are shown in figure 5.19. The growth curves of all cells in the datasets, grouped by cell line, is shown in figure 5.20. The drug panel simulation results can be seen in figure 5.21 and table 5.3. The expression of the top targets in Seurat is displayed in figure 5.22.

Finally, the same was also performed with data from puck three in the Slide-seq spatial RNA-seq dataset (figure 5.23). Differently from the previous simulations, which were run with hexagonal 2D placement, these simulations were run with import placement, which places *microCJ* cells in the same relative location as

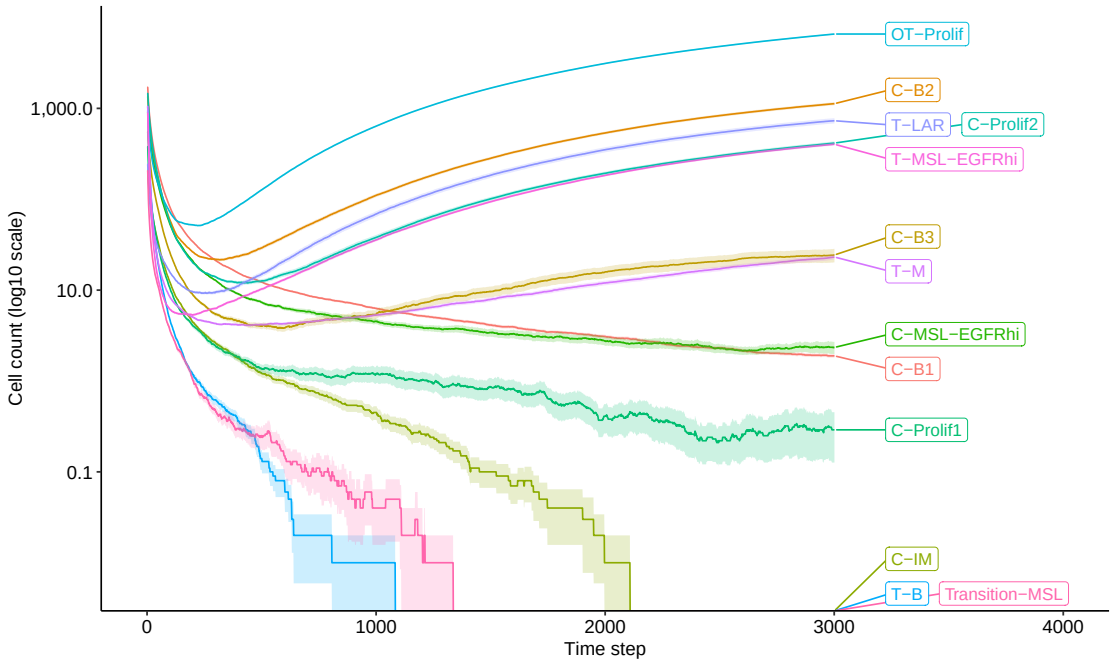


Figure 5.16: Behaviour of the model when fitted with all datasets scRNA-seq data from Savage *et al.*[1]. All populations, imported by single cell from the Savage *et al.*[1] dataset, processed as explained in chapter 2, were simulated in competition for 3000 steps with 100 replicates. The figure shows the average amount of cells per population at each time step, and the standard error of the mean.

the real single cells. Figure 5.24 shows the growth curves for the cells, grouped by cluster. The clusters are the ones identified in chapter 2. One can see that the cluster labelled as mesenchymal/stem-like is the most proliferative by a large margin. The response to the drug panel (figure 5.25, table 5.4) seems to be the same across cell types, although the Basal-like 1 cluster appears to be more sensitive to treatment. The expression of the top targets in Seurat is shown in figure 5.26.

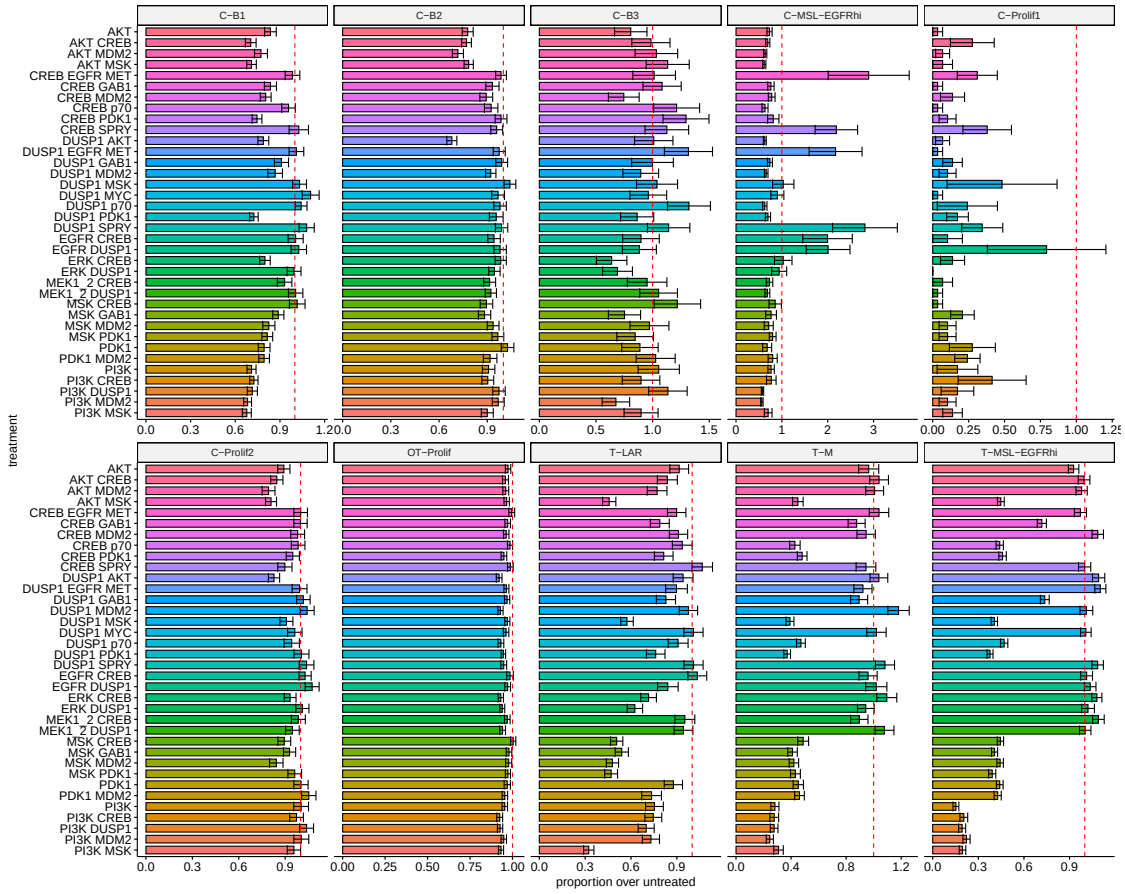
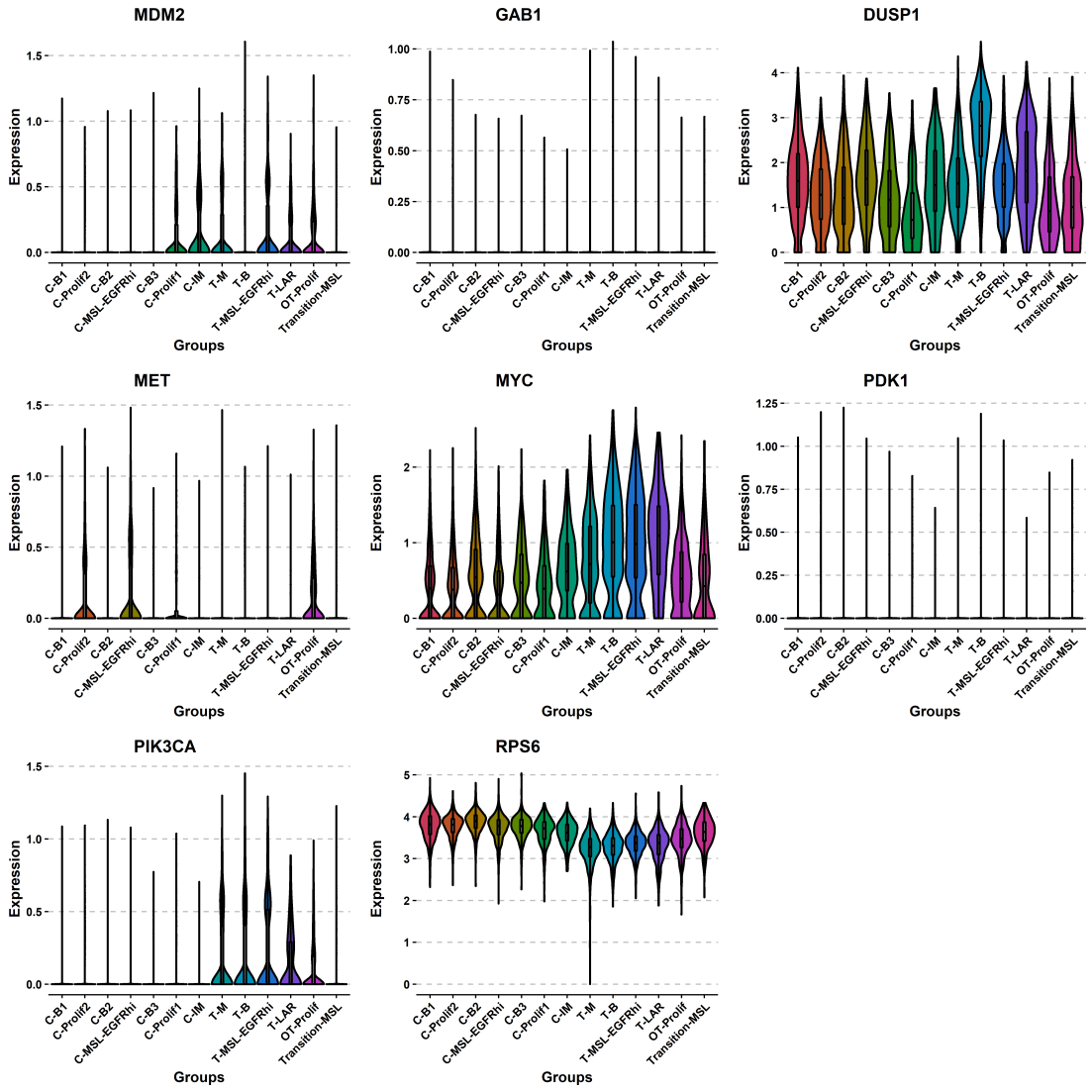
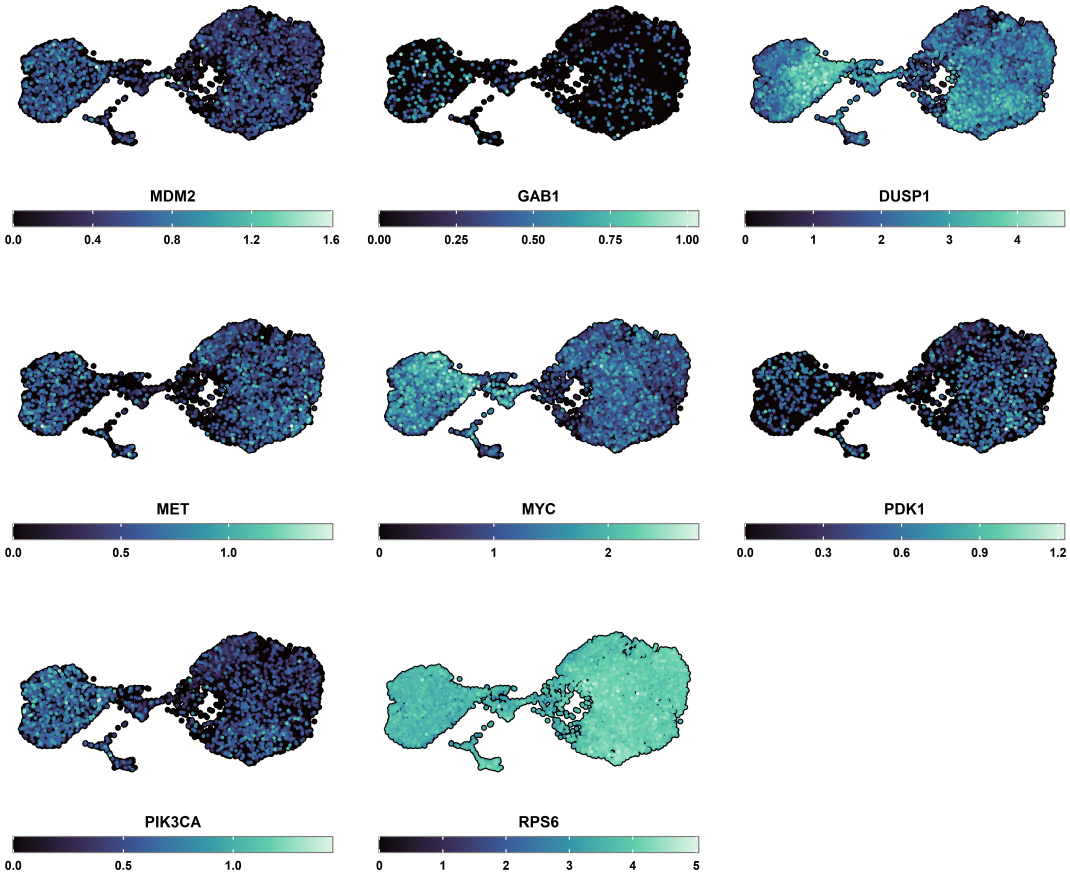


Figure 5.17: Drug panel simulation with Savage *et al.*[1] data. Simulations were run as in figure 5.16, each with a different drug treatment. The figure shows the proportion of alive cells from each subpopulation at the end of the simulation, relative to the untreated simulation (figure 5.16). Each subpopulation is plotted separately. The red dashed line indicates proportion = 1, that is, where the treatment has no effect when compared to the untreated simulation.


MDM2
GAB1
DUSP1
MET
MYC
PDK1
PIK3CA
RPS6
Expression
Groups
C-B1
C-Prolif2
C-B2
C-MSL-EGFRhi
C-B3
C-Prolif1
C-IM
T-M
T-B
T-MSL-EGFRhi
T-LAR
OT-Prolif
Transition-MSL

(a)



(b)

Figure 5.18: Expression of top targets in the Savage *et al.*[1] scRNA-seq dataset. (a) shows a violin plot of each top target across all clusters. In (b) they are represented on a UMAP plot, with cells coloured by gene density (see chapter 2)

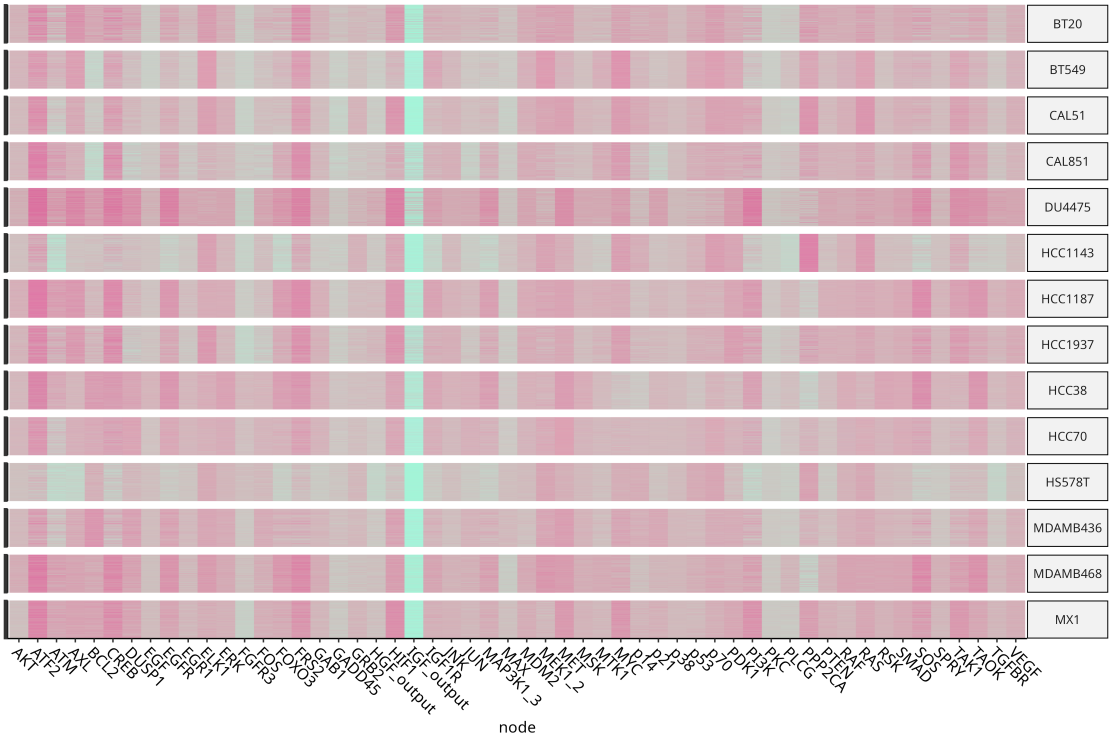


Figure 5.19: An overview of the PBNs produced by *Ingres*, one per each cell, with the dataset from Gambardella *et al.* [297]. The heatmap shows the probabilities for the fixed function (to one or to zero) for each node, as computed by *Ingres*. Each row correspond to a single cell. Cells are grouped by cluster.

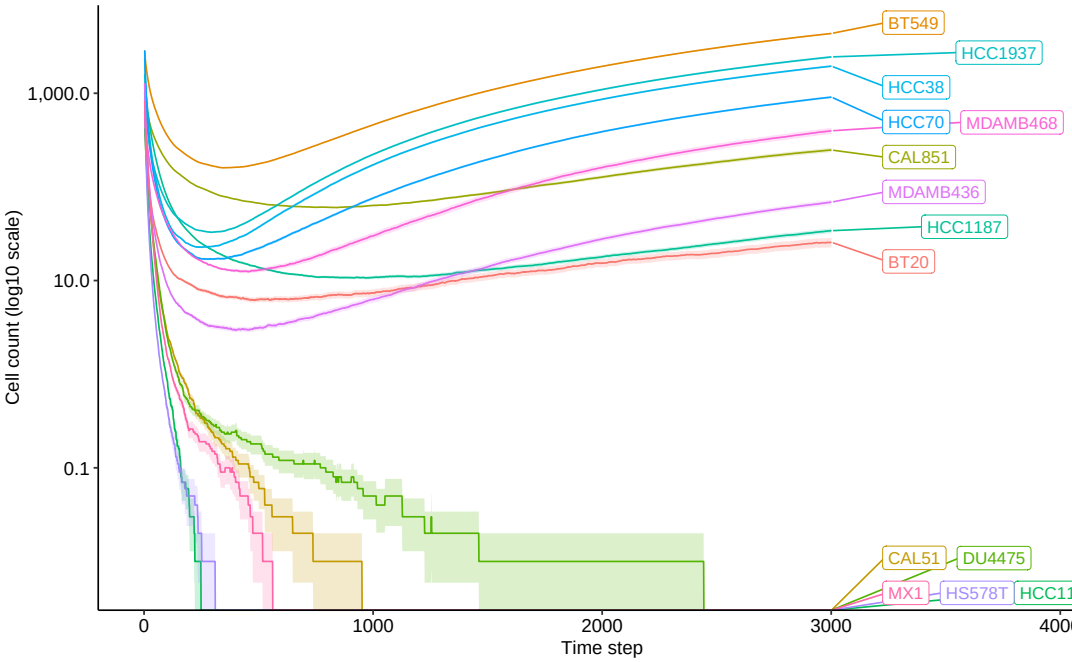


Figure 5.20: Behaviour of the model when fitted with the scRNA-seq data from Gambardella *et al.* [297]. Simulation as in figure 5.16, but with data from Gambardella *et al.* [297], processed as explained in chapter 2.

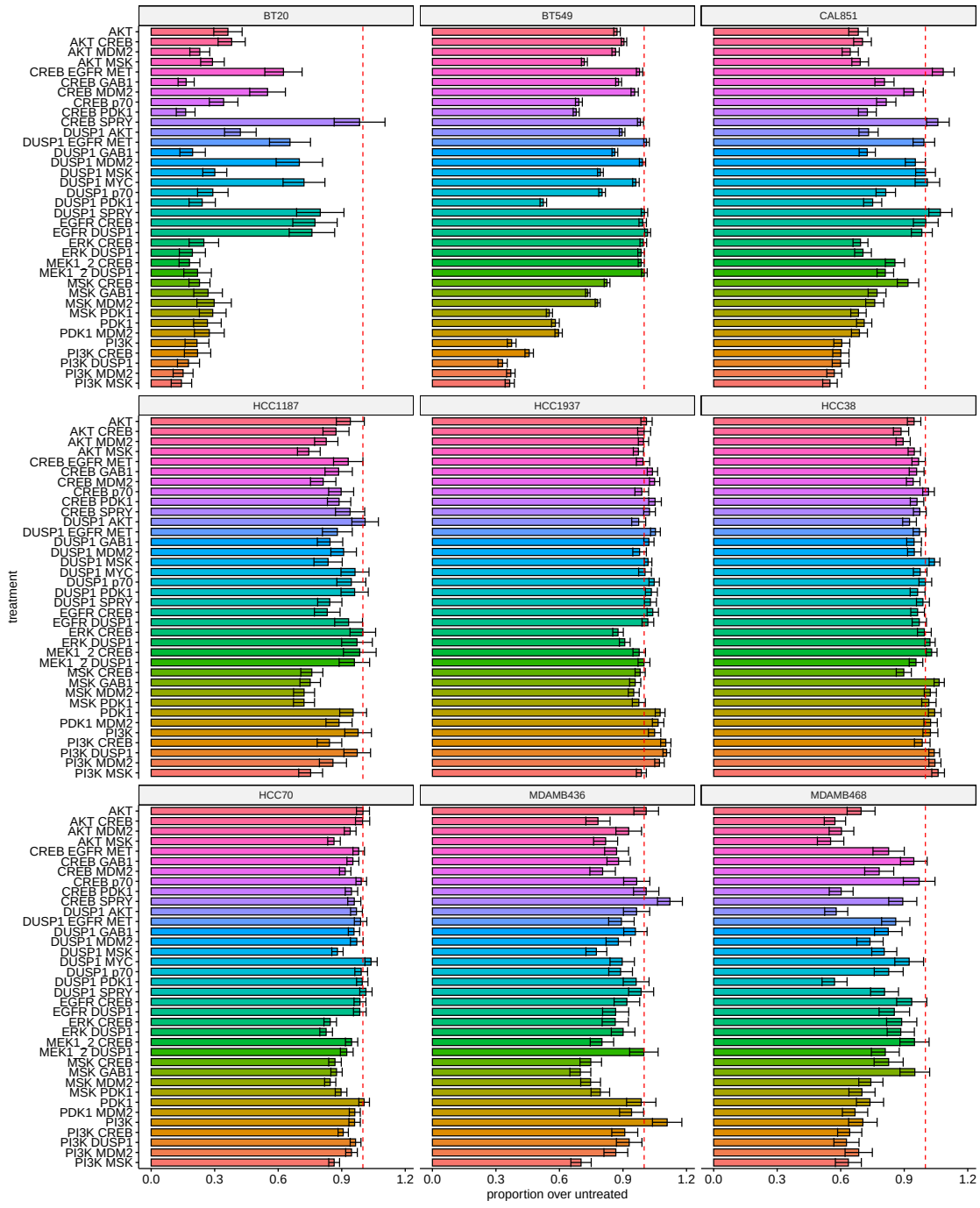
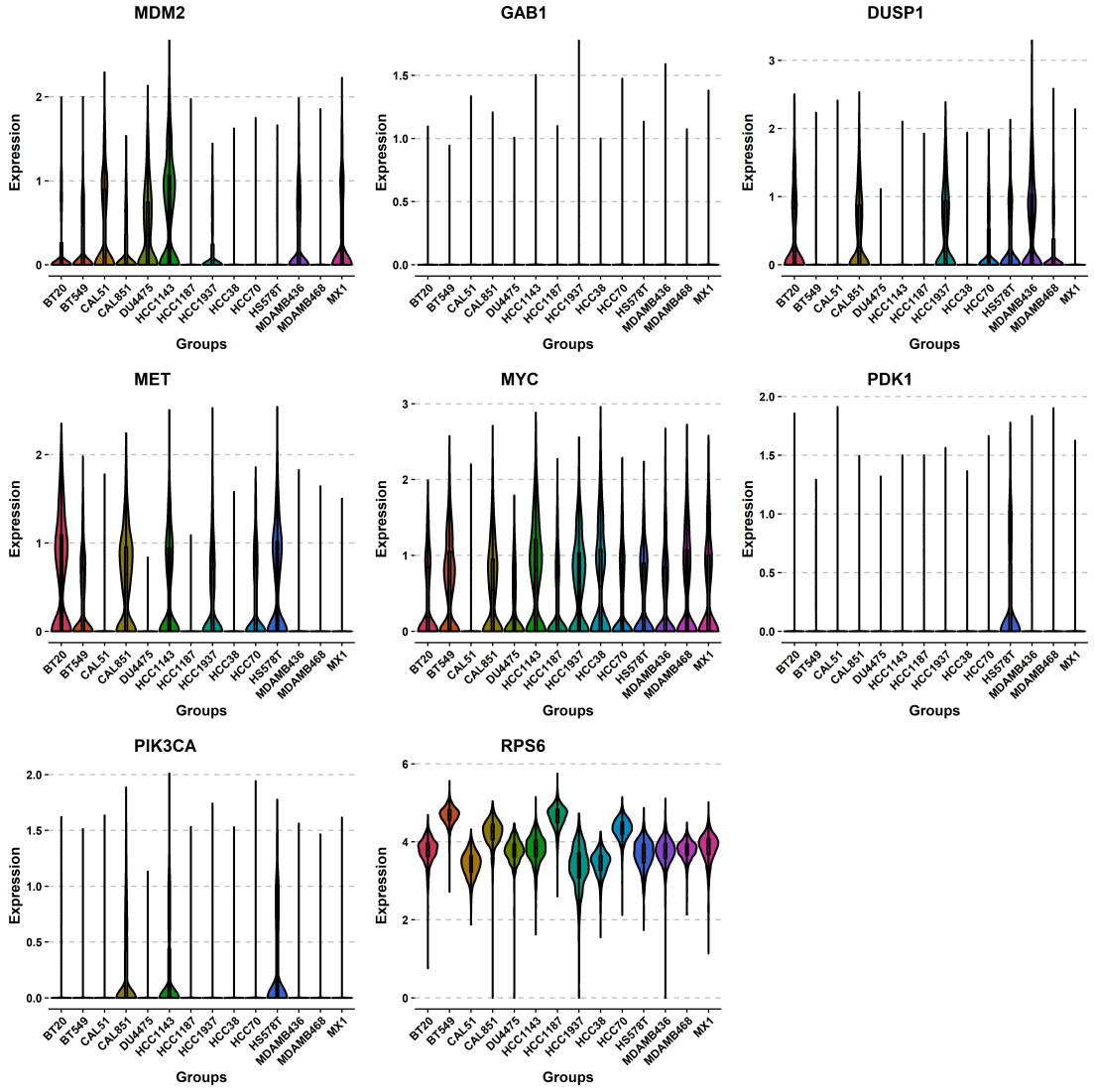
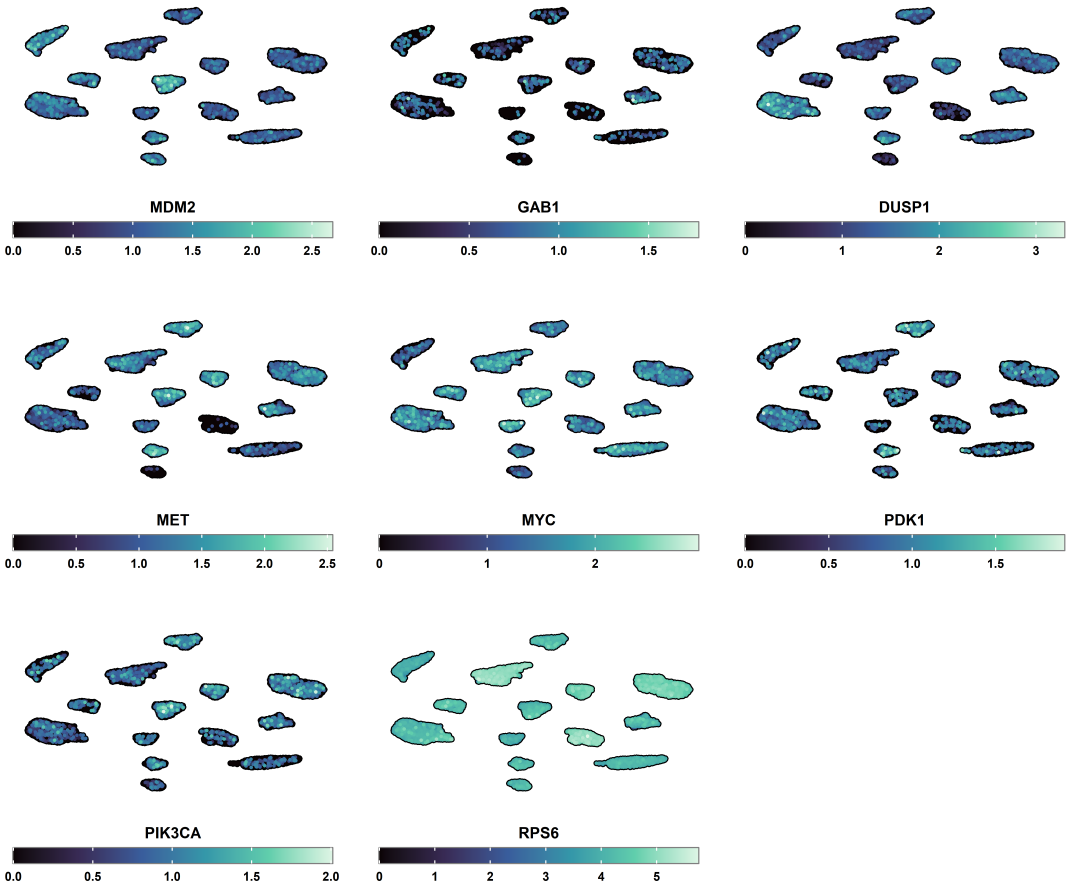


Figure 5.21: Drug panel simulation with Gambardella *et al.*[297] data. Simulation as in figure 5.17, but with data from Gambardella *et al.*[297]. Cell lines that completely died off in the control simulation are not included here.


MDM2
GAB1
DUSP1
MET
MYC
PDK1
PIK3CA
RPS6
Expression
Groups
BT20
BT549
CAL51
CAL851
DU4475
HCC1143
HCC1187
HCC1937
HCC38
HCC70
HS578T
MDAMB436
MDAMB468
MX1

(a)



(b)

Figure 5.22: Expression of top targets in the Gambardella *et al.*[297] scRNA-seq dataset. (a) shows a violin plot of each top target across all clusters. In (b) they are represented on a UMAP plot, with cells coloured by gene density (see chapter 2)

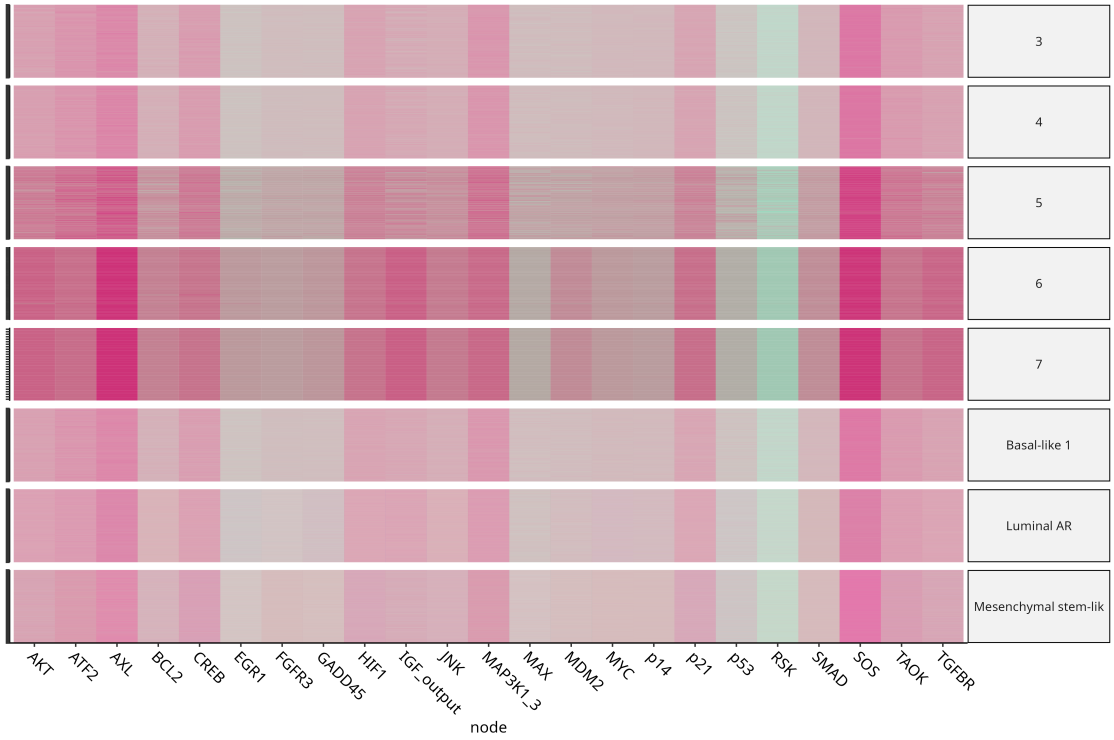


Figure 5.23: An overview of the PBNs produced by *Ingres*, one per each cell, with the Slide-seq spatial RNA-seq dataset (puck 3). The heatmap shows the probabilities for the fixed function (to one or to zero) for each node, as computed by *Ingres*. Each row correspond to a single cell. Cells are grouped by cluster.

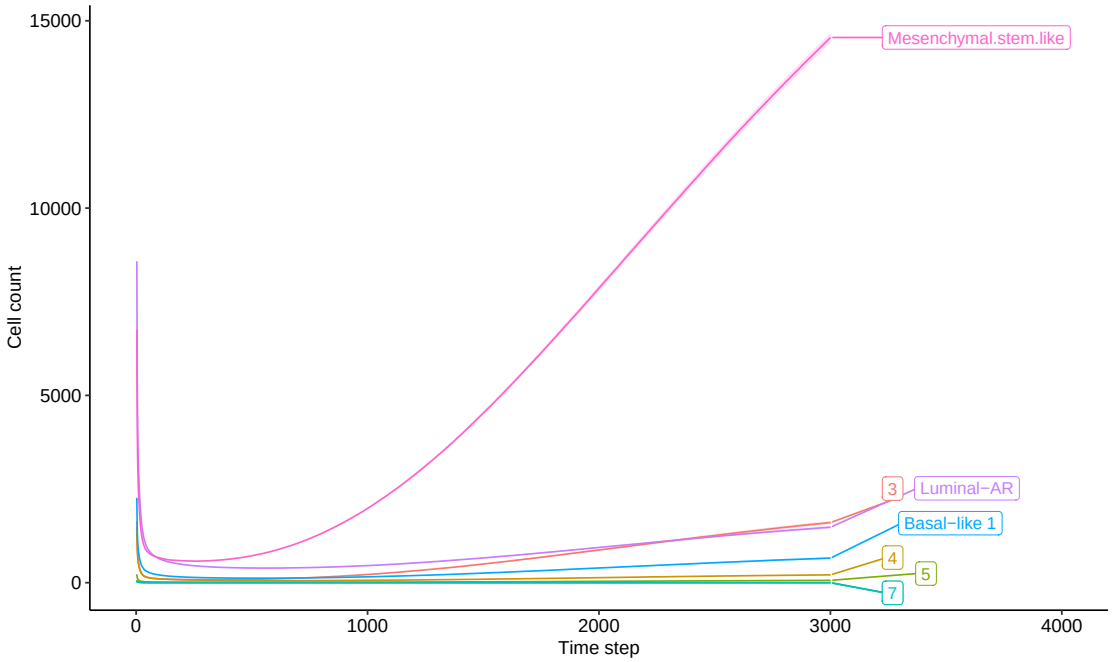


Figure 5.24: Behaviour of the model when fitted with the Slide-seq spatial RNA-seq dataset (puck three). Simulation as in figure 5.16, but with the data from puck three in the Slide-seq spatial RNA-seq dataset, as processed in chapter 2.

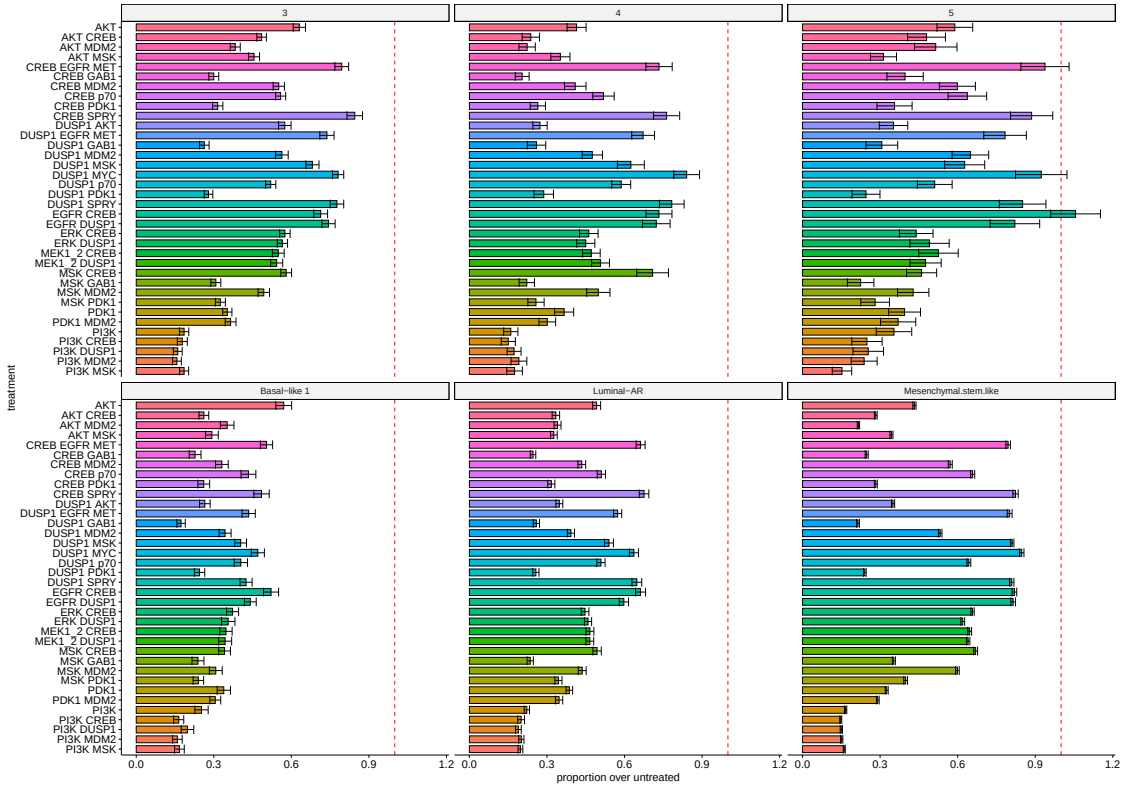
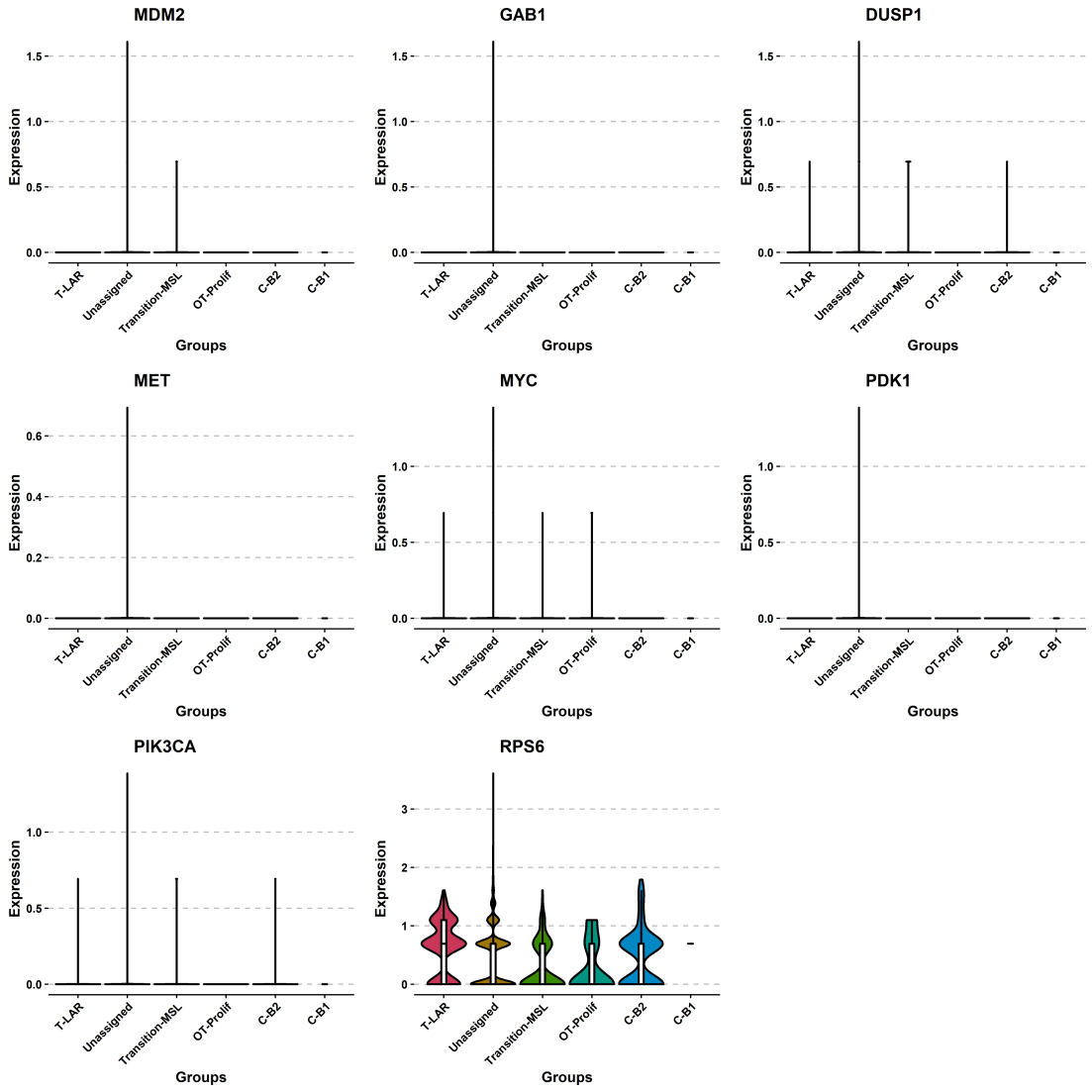


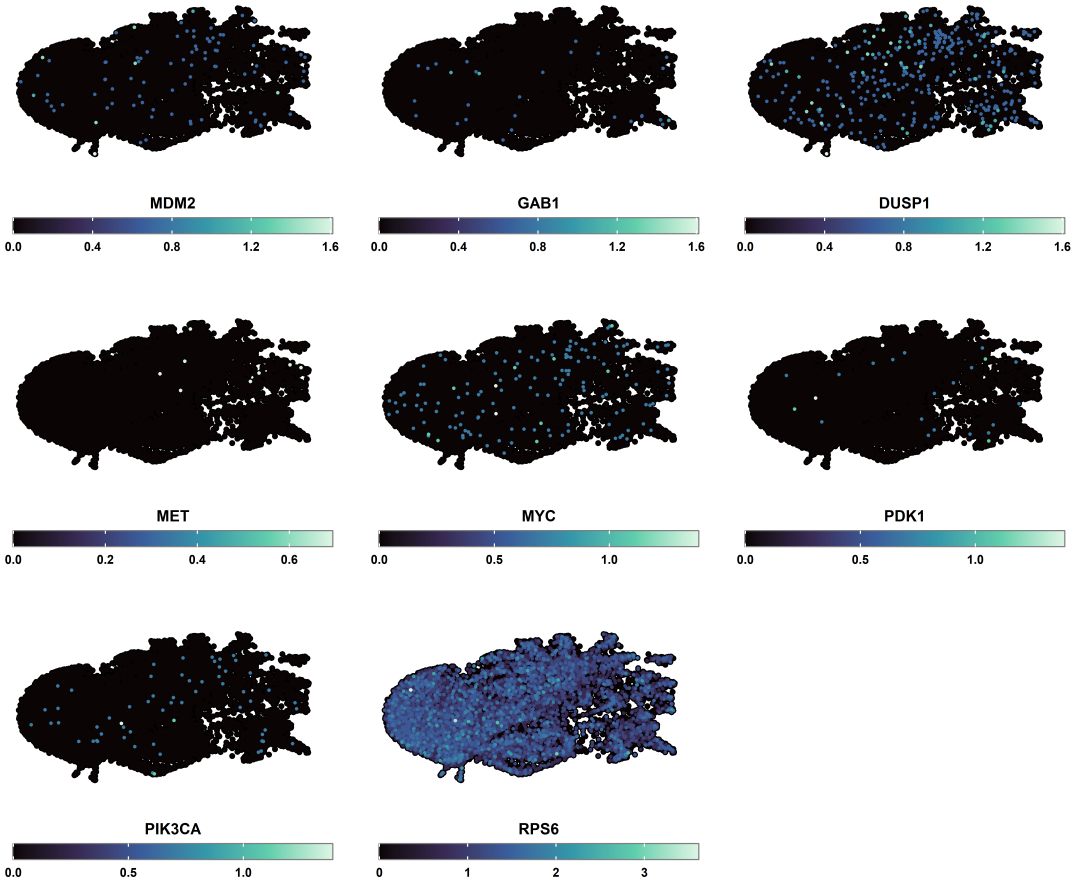
Figure 5.25: Drug panel simulation with the Slide-seq spatial RNA-seq dataset (puck 3). Simulation as in figure 5.17, but with data from puck three in the Slide-seq spatial RNA-seq dataset.

Table 5.4: List of the top 5 most effective drug combinations for each subpopulation, measured as the ones that produce the lowest final average amount of cells, for the Slide-seq dataset.

Drug target	Subpopulations
DUSP1 GAB1	Basal-like 1
DUSP1 PDK1	5
MSK GAB1	5
PI3K	3, 4, Luminal-AR, Mesenchymal.stem.like
PI3K CREB	3, 4, 5, Basal-like 1, Luminal-AR, Mesenchymal.stem.like
PI3K DUSP1	3, 4, Basal-like 1, Luminal-AR, Mesenchymal.stem.like
PI3K MDM2	3, 4, 5, Basal-like 1, Luminal-AR, Mesenchymal.stem.like
PI3K MSK	3, 4, 5, Basal-like 1, Luminal-AR, Mesenchymal.stem.like


MDM2
GAB1
DUSP1
MET
MYC
PDK1
PIK3CA
RPS6
Expression
Groups
T-LAR
Unassigned
Transition-MSL
OT-Prolif
C-B2
C-B1

(a)



(b)

Figure 5.26: Expression of top targets in the Slide-seq spatial RNA-seq dataset. (a) shows a violin plot of each top target across all clusters. In (b) they are represented on a UMAP plot, with cells coloured by gene density (see chapter 2)

5.5 Conclusions

In this chapter I have given a brief overview of the experimental work I performed at the beginning of my DPhil. This was sadly cut short by circumstances provoked by the COVID-19 pandemic. I have also shown how I extended the network used by my model *microCJ* to better capture the EGFR related pathways and possible mechanisms of resistance. This was not very successful, since the drug modelling in *microCJ* needs some refinement and overall, did not produce a model where I could see an effect to RTK targeting drugs. I overcame this by expanding the search for targets to all nodes in the network, which I did by simulating a virtual panel of different drug combinations. I tested the most effective combinations in the models fitted with single-cell and spatial RNA-seq data. *microCJ* ability to import this kind of data is a very promising method for further analysis of this kind of datasets and also to generate new hypotheses and find new possible targets for treatment, as demonstrated in this chapter. Indeed, some possible targets were found. Every combination proposed by the model contained one of AKT, CREB, DUSP1, MSK or PI3K.

AKT and PI3K have long been considered a target for therapy in TNBC. As mentioned in chapter 1, the PI3K/AKT/mTOR pathway is one of the most important pathways associated with cell metabolism, proliferation, differentiation and survival[10]. When deregulated, it plays a crucial role in tumorigenesis and resistance to therapy. The activation of the PI3K/AKT/mTOR pathway is frequent in TNBC, and it occurs through activation of *PIK3CA* or *AKT1* and loss of *PTEN*. This can over-activate the AKT pathway. In general, this pathway presents mutations in up to 25% of breast cancer cases[102]. Several phase II studies have targeted the AKT pathway in TNBC[42,105,106]. In each of them, the combination of an AKT inhibitor, ipatasertib or capivasertib, with the chemotherapy drug paclitaxel resulted in improved PFS and OS. Another phase II trial[345] testing ipatasertib as neoadjuvant therapy in early TNBC was recently finished. The combination of ipatasertib with paclitaxel did not significantly increase the pCR rate, despite a higher overall response rate by MRI. The effect of ipatasertib was

more intense in patients with PIK3CA/AKT1/PTEN altered tumours, highlighting the importance of selecting patients by biomarkers for targeted therapy. PI3K inhibitors are also being tested, such as buparlasib, pictilisib or alpelisib[346] although there are concerns related to increased toxicity[347], which has lead to the early stop of some clinical studies[346].

CREB is the cyclic-AMP response element-binding protein, a nuclear transcription factor that upon phosphorylation binds the transcriptional co-activator, CBP (CREB-binding protein). CREB regulates cell differentiation, proliferation and survival in the nervous system. CREB is involved in tumour initiation, progression and metastasis, and could be considered a proto-oncogene. In the MAPK network I use in *microCJ*, CREB is activated by MSK and activates FOS, DUSP1 and BCL2[313]. CREB is overexpressed in several types of cancer, including breast cancer. Some strategies have been proposed to target it[348], but more research is needed regarding the specificity of treatment and patient selection. Therefore, its selection by *microCJ* as a performant drug in combination with other elements of the PI3K/AKT pathway is interesting, and could merit further experimental testing.

DUSP1 (dual-specificity phosphatase 1), is a negative regulator of the MAPK pathway: it dephosphorylate p38, JNK and ERK[349] — in the network, it inhibits JNK and p38[313]. Despite it often being a tumour suppressor, it can also act as an oncogene, helping tumour cells adapt to high levels of MAPK signaling[349]. Some researchers have look into its potential as a target for inhibition[349,350] in other cancer types, but without clinical data backing it yet. Its presence in the list of top combinations in *microCJ* might derive from the fact that p38 activates p53, and that JNK activates FOXO3, which promotes apoptosis[313]. DUSP1 is a target of CREB, which explains why both are included in the list of top combinations. Nevertheless, this inclusion is intriguing, and like CREB, worthy of experimental validation.

MSK (mitogen- and stress-activated protein kinase), also known as SIK1 (serine/threonine-protein kinase 1), mediates signal transduction downstream of MAPK cascades. It is activated by ERK and p38 and it is ubiquitously expressed in many cell types, but little is know about its specific functions[351]. There is not much

information about its potential as a target in the literature, although a recent article showed that it could be targeted to inhibit medulloblastoma tumorigenesis[352]. In the network, MSK is activated by ERK and p38, as in reality, and activates MYC and CREB, which is the cause for its succesful targeting in *microCJ*. It seems like there are three main axes in the selection of drugs in this panel: AKT/PI3K related targets (GAB1, p70, MYC, and PDK1) RAS downstream genes (ERK and MEK) and CREB related targets (DUSP1 and MSK).

MDM2 (murine double minute 2) is a negative regulator of p53, and as such is an attractive target in TNBC — its inhibition could reactivate p53 and promote tumour suppression. Several studies have been carried out in recent years that show that the targeting of MDM2, by itself or in combination with MDMX (murine double minute X) inhibits *in vitro* and *in vivo* proliferation and invasion. This was confirmed in TNBC cell lines and xenografts when combined with the chemotherapy drug doxorubicin by Fan *et al.*[353]. Therefore, its combination with CREB, MSK or PI3K, as proposed by *microCJ*, seems very attractive and should be followed up in the laboratory.

GAB1 (GRB2-associated binding protein) is implicated in signaling transduction in the EGFR downstream. In the network, it is activated by GRB2 (which is activated by EGFR) and forms a positive feedback loop with PI3K, which makes it clear why GAB1 works as a target in *microCJ*. There is not much data on its use as a target, but Bai *et al.*[354] discovered that microRNA-409-3p, a colorectal tumour suppressor, targets GAB1. This suggests that GAB1 could be an actionable target, but its efficacy and applicability to TNBC is unknown.

p70, also known as RPS6 (ribosomal protein S6), is downstream of the AKT/PI3K/mTOR pathway. It promotes proliferation and cell growth, but its regulation has not yet been thoroughly explored. There are several studies that report that inhibiting or knocking it down have tumour suppressing effects in cancers with high level of p70. Research by Yi *et al.*[137] reported a synergistic downregulation of p70 by targeting EGFR and MET. In the network, p70 directly activates proliferation along with MYC, so it makes sense that targeting it works

well in *microCJ*. Yi *et al.*[137] reported that downregulation of p70 inhibits the proliferation of the MSL TNBC subtype, which is consistent with the results in figure 5.17, where DUSP1 p70 and CREB p70 are most effective in T-MSL-EGFRhi, C-MSL-EGFRhi, but also in T-M and C-Prolif1.

As mentioned before, targeting EGFR and MET did not yield strong results in *microCJ* simulations, despite some preliminary research stating that they are a promising target for combination therapy[137,138,343,344]. However, combining them with a DUSP1 inhibitor seems to have an intense response in the C-Prolif1 subpopulation. They are not in any of the other two top drugs lists. This could indicate that C-Prolif1 mainly relies on overexpression of EGFR, MET and p53 suppression (figure 5.14).

MYC, a proto-oncogene that regulates 15% of all annotated genes, plays essential roles in all cells promoting cell proliferation, growth, apoptosis and differentiation, among other functions. Its deregulation is observed in more than 70% of all human malignancies, and is amplified in 53% of TNBC[355]. This is associated with very poor outcome. Unfortunately, inhibition of MYC remains challenging due to the featureless nature of its protein structure[355], although Wang *et al.*[355] presented a possible strategy to inhibit it using tumour penetrating peptides.

ERK and MEK a downstream effector of RAS signalling, but attempts to target them in the clinical setting have yielded modest results, probably because of feedback activation of alternative pathways. A proposed strategy is to cotarget the RTK responsible for the feedback activation[341].

PDK1 (3-phosphoinositide-dependent kinase 1) plays a crucial role in several cellular processes via the PI3K/AKT dependent or PI3K/AKT-independent pathways in breast cancer cells, such as carcinogenesis, growth and survival, metastasis and chemoresistance. There are several ongoing clinical trials for PDK1-targeted therapy in BC, but no selective inhibitor is currently available, despite the trials showing some benefit to this strategy[356].

Looking at the specific clusters in the Savage dataset, one can see that T-M and T-MSL, which in the drug panel have a good response to PI3K combination

therapy, actually have the highest expression for PI3K and MYC. The same is true for DUSP1 combinations (figures 5.14 and 5.18). Once again, this underlines the importance of selecting patients based on biomarker expression before targeted therapy. It can be seen in figure 5.14 that T-LAR also displays a modestly higher PI3K expression (figure 5.18 a) which is translated into some sensitivity to PI3K combination therapy, but not as much as for T-M and T-MSL-EGFRhi. This could mean that the mechanism of resistance to EGFR targeted treatment in this dataset is that the surviving cells overactivate parallel pathways that end up overexpressing the downstream of the PI3K/AKT pathway, as well as MYC. This method of resistance is one of the most common for EGFR targeted drugs, and it has been extensively discussed in the literature[340,341]. This of course would have to be validated experimentally, but it highlights the usefulness of computational models to make sense of multidimensional data and highly complex pathways.

Another interesting fact is that while C-Prolif1, C-Prolif2 and OT-Prolif cells have very similar *Ingres* PBNs (figure 5.14), OT-Prolif cells perform much better in the simulations, with or without treatment (figures 5.16 and 5.17). But if one looks closely at the OT-Prolif profile (figure 5.14), it is more heterogeneous than the other two, with the bottom rows (cells) having NES of the opposite sign to the rest (p53, for example, is mostly negative for this cells, whereas it is positive for the rest). I hypothesise that those cells are the very few ones that survive after a few iterations of the simulation, and then take over most of the simulated tumour. Again, this points at the fact that heterogeneity is one of the main causes for drug resistance and tumour survival.

Turning to the TNBC results (figure 5.21) one can see something similar, the cell lines that are sensitive to PI3K targeting have a slightly higher PI3K expression and *Ingres* NES (figures 5.19 and 5.22), but this is not as marked as for the Savage dataset and it only occurs in some of the cells (figure 5.19). This is consistent with one of the authors' point, that cell lines also exhibit heterogeneity and differential response to targeted treatment.

Of note, the most proliferative cell line in the simulation, BT549, is also highly sensitive to PI3K combination therapy, which was not the case for the most proliferative subpopulations in the Savage *et al.* [1] dataset.

Two of the most PI3Ki and AKTi sensitive cell lines, BT20 and MDA-MB-468, present a high EGFR expression (figure 2.20, which of course was also true for the treatment clusters in the Savage *et al.* [1] dataset. Looking at each cell line molecular profile[357] BT20 presents EGFR amplification and p53 and PI3K oncogenic mutations. BT549 displays a p53 oncogenic mutation and PTEN deletion. HCC1937 has BRCA and p53 oncogenic mutations, and PTEN deletion. MDA-MB-436 presents BRCA and p53 oncogenic mutations. Finally, MDA-MB-468 displays EGFR amplification, PTEN deletion and a p53 oncogenic mutation. Taking all this information together, it seems that cell lines with EGFR amplification are more sensitive to AKTi and PI3Ki combination therapy. In fact, it has been reported[347] that cases with PI3K mutation or PTEN loss are potential targets for allosteric inhibitors of AKT. On the other hand, there is not a clear marker for PI3K sensitivity in this model — only BT20 has a PI3K mutation, but it is not the only one responding to the therapy, however it is the most sensitive. PI3K mutation then confers certain predictive value for response to PI3Ki combination therapy, but not enough. This needs to be completed with more refined models and experimental work.

As for the spatial dataset simulations, it can be seen that many nodes in the network were not found by *Ingres* (figure 5.23), likely due to the higher number of gene dropouts in this kind of data compared to non-spatial RNA-seq datasets (which is already high). This, together with the realisation that all clusters have a similar response to treatment (figure 5.25), leads me to not deem these results trustworthy, and it points out one of the limitations in this pipeline: with low quality data, the model will not work properly, since the biases imposed on its networks (figure 5.23) will be quite extreme and homogeneous, not bearing much resemblance to reality, and producing a very monolithic model that fails to capture the heterogeneity and capability for adaptation of real tumours. However, one of the subpopulations, the mesenchymal/stem-like, has a significantly higher proliferative

rate than the others, and just by looking at the *Ingres* heatmap (figure 5.23) it is not instantly clear why that is. Again, some heterogeneity may be present, and just a few cells are being selected but, not encountering much competition, are proliferating as much as possible.

Despite the sensible results shown in this chapter, this pipeline still needs more development. One problem is, as mentioned above, that only the cells with the most extreme enrichment scores superimposed on their networks survive throughout the simulation. This causes only the more rigidly simulated cells — those whose behaviour is more dependent on the static enrichment scores derived from *Ingres* than on the dynamic behaviour of their network — are selected. Ideally, not only cells who are proliferating at the highest rate should survive. Instead, in a more refined version of the model, one might see different phenotypes being displayed: instead of only growth arrest, apoptosis and proliferation, one might also model quiescence, senescence, and cell migration. A related problem is that the predictions of the model can only be as detailed or unbiased as the Boolean network in the cells. The network I have used in this project is quite small when compared to all the possible targets and pathways of resistance in TNBC. On the other hand, for models to be useful they have to discard information, otherwise they risk being overfitted and too static, on top of prohibitively computationally expensive. Other ways to bring together gene networks and transcriptomics data could also be explored, preferably ones with more range for dynamic behaviour, such as qualitative networks[358]. *microCJ* and *Ingres* are released as open-source, copyleft software, so I expect other researchers in the field might contribute and expand on the basis I have developed, to add more possible phenotypes and more refined experimental data import pipelines. All in all, I think this chapter is a good demonstration of the ideal feedback loop between experimental data and computational modelling, how they can both inform each other, and the kind of results this way of working can yield.

5.6 Methods

5.6.1 *In vitro* methods

5.6.1.1 Western blot

Cells were plated in 60 mm culture dishes at a density of $2 \cdot 10^5$ per dish, in low glucose DMEM media supplemented with FBS 10%. Cells were incubated for 24 hours in normoxic conditions. After this, cells were treated with gefitinib 10μM dissolved in DMSO, and the equivalent amount of DMSO was added to the control samples. Half of the samples were then moved to a hypoxia chamber where they were incubated at an oxygen concentration of 1%. The rest were put back in a normoxic incubator. After 24 or 48 hours, depending of the sample, cells were harvested, washed with cold PBS and lysed using a lysis buffer containing RIPA buffer, 1X phosphatase inhibitor (PhosSTOP, Roche) and 1X protease inhibitor (Complete mini, Sigma Aldrich). The lysates were vortexed and centrifuged at 12000 rpm for 20 minutes at 4°C. The supernatant was collected and protein concentration was determined using Bio-Rad protein assay. Samples containing equal amounts of proteins (20μg) were separated on Tris-Acetate SDS-PAGE gels (ThermoFisher) and transferred to PVDF membranes (Millipore). Membranes were blocked for 20 minutes with 5% BSA in 1X TBS 0.1% Tween at room temperature. Then, membranes were incubated in anti-phospho-EGFR antibody Tyr1068 (Cell Signaling) at a final dilution of 1:500 and anti-β-actin (Abcam) at 1:1000, overnight at 4°C. After five washes with TBS-Tween, membranes were incubated with secondary antibodies for an hour, goat anti-mouse IgG-HRP for β-actin and goat anti-rabbit for pEGFR (ThermoFisher). After five more washes, detection was performed using ECL chemoluminescence reagents (Life Sciences). Then, after five washes with TBS-Tween, membranes were washed with Restore stripping buffer (ThermoFisher) for 15 minuter at room temperature. After blocking with BSA again, membranes were incubated with anti-EGFR antibody (Cell Signaling), followed by the same protocol described above.

5.6.1.2 Flow cytometry

A million cells were centrifuged and resuspended in cold PBS containing Live/Dead fixable FarRed dye (ThermoFisher). Then the conjugated PE anti-EGFR antibody (BD BioSciences) was added at 1:10 concentration, with an incubation of 10 minutes at 4°C in the dark. Cells were then washed with PBS and centrifuged at 1800 rpm for two minutes. The resulting pellet was resuspended in PBS and run in a FACSCalibur machine. A PE isotype (BD BioSciences) was used as negative control. Analysis of the results and gating were performed using the FlowJo software.

5.6.1.3 Proliferation assays

Cells were seeded into 96-well plates in triplicates, at a density of 10,000 cells/well, in low glucose DMEM 10% FBS. Cell were incubated for 24 hours in normoxia. Then, cells were treated with increasing concentrations of gefitinib and half of the samples was moved to a hypoxia chamber at 1% oxygen. After 24, 48 or 72 hours, depending on the sample, plates were collected and frozen at -80°C. Then, cell number was measured using CyQUANT assay (ThermoFisher).

5.6.2 Importing of transcriptomics data

All single cell datasets were re-normalised and re-scaled and centered in Seurat using the functions `NormalizeData()`, `FindVariableFeatures()` and `ScaleData()` before importing into *Ingres*.

6 Discussion

In this thesis I have presented the computational model of cancer I developed during my DPhil. *microCJ* is a multiscale, agent-based model, where intracellular and intercellular interactions are simulated. A gene regulatory Boolean network is independently run inside each cell in the simulation. These networks govern the behaviour of each cell: whether it proliferates or dies. Researchers can introduce perturbation in these networks as mutations or different levels of expression, or the cells can autonomously acquire new mutations during the course of the simulation. At the intercellular level, cells compete with each other for space to grow in and nutrients. A diffusion simulation is carried out in unison with the cell model, where substances diffuse through the simulation space and concentration gradients of oxygen, growth factor or drugs are formed. Importantly, drug diffusion forms the basis for resistance to therapy simulations.

The model has many more features that have been described in chapter 4. In that chapter I also went through the extensive testing I performed for each feature, and the rationale behind each modelling decision. I hope *microCJ* will be useful for any scientist that want to employ computational modelling to aid in their understanding of tumour heterogeneity, of how specific driver mutations give rise to the different phenotypes observed in tumours, to explain and test hypotheses and in summary, to help make sense of this extremely complex and deadly disease.

microCJ will be released as a free, open-source application that other researchers can use, extend and adapt to their own research focus and needs. I designed it with extensibility and modularity in mind, and every part of the simulation can be configured and easily modified, which I think is one of the greatest advantages of my model as a tool for other researchers.

microCJ was inspired by microC[257], but adds many new features, such as the ability to simulate asynchronous and probabilistic Boolean networks, probabilistic mutations and cells acquiring new mutations during the simulation, different algorithms for the placement of cells both on-lattice and off-lattice, the capacity of cells to push others to make space for new growth, or temporary drug administration schedules. The implementation is also easier to extend due to its modularity and being coded in a mainstream programming language, Java.

I then used this model to explore TNBC, a highly aggressive, with very poor prognoses, type of cancer with an unmet need. To do that, I employed several RNA-seq datasets, obtained from collaborators. These datasets help paint a picture of TNBC as a highly heterogeneous, multiform disease. I analysed these data in chapter 2. Of particular interest was the work of Savage *et al.*[1], who examined an exceptional responder to EGFR-targeted therapy in TNBC. EGFR is an interesting protein in this cancer type. It is usually overexpressed in TNBC, but attempts to use it as a successful target for therapy have so far been disappointing.

In order to integrate these RNA-seq data into *microCJ*, I devised a tool to fit the Boolean networks that cells in the model run with single-cell RNA-seq data. This tool, *Ingres* (described in chapter 3) creates a probabilistic Boolean network for each cell in the data. These PBNs are fitted with a probability to be active or inactive relative to the protein activity enrichment score computed by the VIPER algorithm. This way, each PBN will represent the overexpression or underexpression of each node in the network as a given bias towards activation or inactivation that is much more granular than the binary state in Boolean nodes.

Using these tools together, I simulated TNBC growth in *microCJ*, fitted with the different RNA-seq datasets I had previously analysed. *microCJ* maintained,

to a certain degree, the basic phenotypes of each cluster in the data, that is, cell proliferation or death. Due to the simplicity of the network the model currently uses, it failed to offer explanations for resistance to EGFR targetting, other that the ones already known in the literature and that I could fit into the network. The next steps for this would be to greatly extend the network to offer a more complete representation of the current biological knowledge of the TNBC landscape. This way, the model could really offer explanations that are not biased by the specific focuses of the network, in this case, the MAPK pathway. This should be done carefully, because every extension of such a network treads a fine line between completeness and overfitting. The more pathways or interactions are added to the network, the more salient the things one chooses to leave out become. And while specificity should not be a problem, since it is expected that a different network is used for every particular research project, the model should not become so complex that it becomes intractable or mechanistic assumptions start to become fuzzy.

Drug simulation is another area that needs to be improved in *microCJ*, to properly model drug response. This should be done taking extensive data about *in vitro* and *in vivo* responses to an specific drug, and model its timescale of activity, its specificity and effectiveness. Only then the model could be expected to accurately (as much as possible) predict the response of a tumour to not yet tested drug combinations. And of course, any predictions made by the model should be tested experimentally.

Despite these hurdles, I opted to test the TNBC fitted *microCJ* with a panel of drugs. This offered some interesting results, namely the efficacy of uncommon or poorly understood targets such as CREB, DUSP1, MSK, MDM2 or GAB1. Top drug combinations in the model frequently included PI3K and AKT, which have been the subject of many clinical trials with variable degrees of success, with AKT being the most promising. Notwithstanding the aforementioned difficulties with drug modelling, the targets proposed in chapter 5 merit further experimental research. Originally, I planned to perform some of this experimental validation myself, but difficulties caused by the global circumstances made that part of my project impossible. Nevertheless, I hope that my work with *microCJ* proves that

such computational models are a useful tool to point at new targets for therapy, that can then be looked at in the laboratory. Indeed, one of the main advantages of modelling is that, while drug panels are expensive and time-consuming, a model such as *microCJ* can test dozens or hundreds of drug combinations, serving as a first filter and test that informs which candidate drugs should be examined experimentally.

My work on *microCJ* was done with TNBC in mind, but the truth is that this model can be used with any kind of cancer. For future development, there are several features I think should be added to *microCJ*, apart for the already mentioned above. First, methods for automated building and fitting of networks using transcriptomics data should be explored, so that networks with hundreds of nodes would be created, fitted so that the best explanation to the available data is reached. Some work has been done towards this, such as CARNIVAL[187].

The microenvironment and the tumour interaction with it is increasingly being recognised as an crucial element in cancer biology. So far, *microCJ* simulation of the microenvironment only includes the cell-cell exchange of signals and growth inhibition, but the interactions with the extra-cellular matrix could also be modelled. Another essential area in oncology is immunology. Immune cells could be modelled into *microCJ* as separate agents with their own, immune-specific networks, and their interactions with cancer cells. This would be the basis for a model of immune therapy.

Finally, one of the main limitations of *microCJ* is that only considers a few possible phenotypes for the cells, mainly proliferation or apoptosis. These could be extended to include behaviours such as quiescence and senescence, cell migration and metastasis, and different stages and outcomes of the cell cycle. Another related feature that should be added is for cells to display several phenotypic features at once, such as migrating growing cells.

I do not envision all of this features being added at once to the model, but rather, continuing with the modular basis I have developed, a series of modules that can be included or excluded as necessary for each specific purpose. As previously stated, *microCJ* flexibility is its main asset, and I look forward to future additions by other researchers with excitement.

Appendices

A

Documentation and tutorial vignette for the *Ingres* package

Using Ingres to obtain single-cell probabilistic Boolean networks from single-cell RNA-seq data

Overview of Ingres

Gene regulatory networks (GRN) have long been used as a way to integrate heterogeneous data into a discrete model, and are very useful to generate actionable hypotheses on the mechanisms governing these biological processes. Here we present Ingres (Inferring Probabilistic Boolean Networks of Gene Regulation Using Protein Activity Enrichment Scores) an open-source tool that uses
single-cell sequencing data and prior knowledge GRNs to produce a probabilistic
Boolean network (PBN) per each cell and/or cluster of cells in the dataset. Ingres allows to better capture the differences between cell phenotypes, using a continuous measure of protein activity while still confined to the simplicity of a GRN. We believe Ingres will be useful to better understand the heterogeneous makeup of cell populations, to gain insight into the specific circuits that drive certain phenotypes, and to use expression and other omics to infer computational cellular models in bulk or single-cell data.

Specifically, we use a previously developed algorithm, VIPER[1], to infer protein activity starting from a gene expression matrix and a list of regulons —defined as the list of transcriptional targets of each protein. This computes a matrix of normalised enrichment scores (NES) for each protein. The expression of these regulons represent an optimal indirect method to measure the activity of a specific gene/protein (the nodes in the preexisting Boolean network). This constitutes a novel way to fit the PBN using single-cell RNAseq data.

For more details on the specific algorithms used, please refer to our publication describing Ingres[2].

Citation

Victori, P. & Buffa, F. M. Ingres: from single-cell RNA-seq data to single-cell probabilistic Boolean networks. 2022.09.04.506528 Preprint at https://doi.org/10.1101/2022.09.04.506528 (2022).

Dependencies

Most of Ingres dependencies are available on CRAN. Some of them need to be installed from Bioconductor:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("viper")
BiocManager::install("org.Hs.eg.db")
BiocManager::install("AnnotationDbi")
BiocManager::install("aracne.networks") # optional dependency
```

Creating the Ingres object

First we load the Ingres package:

```
library(ingres)
```

Then we need to create an Ingres object that will hold all the necessary data. Ingres contains example data to show how to use its functions, we will print them in this vignette so their structure can be seen. There are two main ways of creating an Ingres object. The easiest one is using a `Seurat`[3] object with the scaling and clustering pipelines already performed (see the corresponding introductory vignette to Seurat):

```
data(small_blca_wang)
small_blca_wang
#> An object of class Seurat
#> 16314 features across 90 samples within 1 assay
#> Active assay: RNA (16314 features, 2000 variable features)
#>  1 dimensional reduction calculated: tsne
```

We also need a network object in `tidygraph` format:

```
data("network")
network
#> # A tbl_graph: 30 nodes and 110 edges
#> #
#> # A directed multigraph with 1 component
```

```
#> #
#> # Node Data: 30 × 3 (active)
#>   kind  rule                              id
#>   <chr> <chr>                             <chr>
#> 1 input ""                                DNAdamage
#> 2 input ""                                GrowthInhibitors
#> 3 fate  "CyclinE1 or CyclinA"             Proliferation
#> 4 fate  "E2F1_2 or TP53"                  Apoptosis
#> 5 fate  "p21CIP or RB1 or RBL2"           Growth_Arrest
#> 6 gene  "SPRY and not FGFR3 and not GRB2" EGFR
#> # … with 24 more rows
#> #
#> # Edge Data: 110 × 3
#>    from    to sign
#>   <int> <int> <chr>
#> 1     8    30 positive
#> 2    14     3 positive
#> 3    15     3 positive
#> # … with 107 more rows
```

Ingres includes several helper functions that can transform GinSim or `graphml` files to the `tidygraph` format.

```
# Load the graphml example file
filename = system.file("extdata", "example_network.graphml", package = "ingres")

#Convert to tidygraph format
graphmlAsTidy(filename)
#> # A tbl_graph: 30 nodes and 110 edges
#> #
#> # A directed multigraph with 1 component
#> #
#> # Node Data: 30 × 3 (active)
#>   kind  rule                              id
#>   <chr> <chr>                             <chr>
#> 1 input ""                                DNAdamage
#> 2 input ""                                GrowthInhibitors
#> 3 fate  "CyclinE1 or CyclinA"             Proliferation
#> 4 fate  "E2F1_2 or TP53"                  Apoptosis
#> 5 fate  "p21CIP or RB1 or RBL2"           Growth_Arrest
#> 6 gene  "SPRY and not FGFR3 and not GRB2" EGFR
#> # … with 24 more rows
#> #
#> # Edge Data: 110 × 3
#>    from    to sign
#>   <int> <int> <chr>
#> 1     8    30 positive
#> 2    14     3 positive
#> 3    15     3 positive
#> # … with 107 more rows
```

```
# Load the GinSim (.zginml) example file
filename = system.file("extdata", "example_ginsim.zginml", package = "ingres")

# Convert to graphml and store it in a temporary file
temp = tempfile()
gml = ginmlToGraphml(filename, dest = temp)
head(gml)
#> [1] "<?xml version=\"1.0\" encoding=\"UTF-8\"?><graphml
       xmlns=\"http://graphml.graphdrawing.org/xmlns\">"
#> [2] "<key attr.name=\"kind\" attr.type=\"string\" for=\"node\" id=\"kind\"/>"
#> [3] "<key attr.name=\"rule\" attr.type=\"string\" for=\"node\" id=\"rule\"/>"
#> [4] "<key attr.name=\"sign\" attr.type=\"string\" for=\"edge\" id=\"sign\"/>"
#> [5] "<graph edgedefault=\"directed\">"
#> [6] "<node id=\"DNA_Damage\">"

# Convert to tidygraph
graphmlAsTidy(temp)
#> # A tbl_graph: 26 nodes and 71 edges
#> #
#> # A directed simple graph with 1 component
#> #
#> # Node Data: 26 × 3 (active)
#>   kind  rule                                              id
#>   <chr> <chr>                                             <chr>
#> 1 input ""                                                DNA_Damage
#> 2 gene  "DNA_Damage and (E2F1 or not Wip1 or not HDAC1)" ATM
#> 3 gene  "ATM and not Wip1"                                p38MAPK
```

```
#> 4 gene  "(not Wip1 or p53) and not ATM"                Mdm2
#> 5 gene  "ATM or (not Mdm2 and p38MAPK)"                p53
#> 6 gene  "p53_A"                                        Wip1
#> # … with 20 more rows
#> #
#> # Edge Data: 71 × 3
#>    from    to sign
#>   <int> <int> <chr>
#> 1    18     8 negative
#> 2    18     7 negative
#> 3    18    17 negative
#> # … with 68 more rows
```

A data frame with a row per node in the network and its corresponding gene symbol, in case nodes in the network have different names:

```
data("network_genes")
network_genes
#>        node   symbol
#> 1      EGFR     EGFR
#> 2     FGFR3    FGFR4
#> 3       RAS     HRAS
#> 4       RAS     KRAS
#> 5       RAS     NRAS
#> 6      E2F1     E2F1
#> 7  CyclinD1    CCND1
#> 8  CyclinE1    CCNE1
#> 9   CyclinA    CCNA1
#> 10  CyclinA    CCNA2
#> 11   CDC25A   CDC25A
#> 12 p16INK4a p16INK4a
#> 13      RB1      RB1
#> 14     RBL2     RBL2
#> 15   p21CIP   CDKN1A
#> 16      ATM      ATM
#> 17  CHEK1_2    CHEK1
#> 18  CHEK1_2    CHEK1
#> 19     MDM2     MDM2
#> 20     TP53     TP53
#> 21   p14ARF   CDKN2A
#> 22     PTEN     PTEN
#> 23     PI3K   PIK3CA
#> 24      AKT     AKT1
#> 25      AKT     AKT2
#> 26      AKT     AKT3
#> 27     GRB2     GRB2
#> 28     SPRY    SPRY2
```

We can use the function `createNetworkGenesTemplate` to generate a network_genes file with the node names prepopulated. If using RStudio interactively, the function opens this file for us to modify as needed.

```
# store and modify = FALSE just for demonstration.
createNetworkGenesTemplate(network, store = FALSE, modify = FALSE)
#> # A tibble: 25 × 2
#>    node     symbol
#>    <chr>    <chr>
#>  1 EGFR     EGFR
#>  2 FGFR3    FGFR3
#>  3 RAS      RAS
#>  4 E2F1_1   E2F1_1
#>  5 E2F1_2   E2F1_2
#>  6 E2F3_1   E2F3_1
#>  7 E2F3_2   E2F3_2
#>  8 CyclinD1 CyclinD1
#>  9 CyclinE1 CyclinE1
#> 10 CyclinA  CyclinA
#> # … with 15 more rows
# Then, modify as needed.
```

Then, we can create the Ingres object:

```
ing = createIngresObjectFromSeurat(
  seurat.object = small_blca_wang,
  seurat.assay  = "RNA",
  slot          = "data",
  network.genes = network_genes,
  network       = network
```

```
)

ing
#> An ingres object
#> --------------------
#> 16314 features across 90 samples
#> 9 clusters
#> Its Boolean network has 30 nodes
#> Viper not performed
```

An Ingres object can also be created without a `Seurat` object, by providing a expression matrix, with cells in rows and genes in columns, and a idents data.frame with each cell barcode and its corresponding cluster or subpopulation:

```
exp = ing@expression
exp[1:2, 1:2]
#>              sample1@AGCTTGAGTTCCTCCA-1 sample1@CAGCAGCCAGTCTTCC-1
#> RP11-34P13.7                          0                           0
#> AL627309.1                            0                           0

idents = ing@idents
head(idents)
#>                           cell cluster
#> 1 sample1@AGCTTGAGTTCCTCCA-1       5
#> 2 sample1@CAGCAGCCAGTCTTCC-1       5
#> 3 sample1@CTAGCCTGTTAGTGGG-1       5
#> 4 sample1@GATGAGGGTCTGGTCG-1       5
#> 5 sample1@TGACTTTTCAAGAAGT-1       5
#> 6 sample1@TGCGTGGGTGTTGGGA-1       5
```

```
createIngresObject(
  expression.matrix = exp,
  idents            = idents,
  network.genes     = network_genes,
  network           = network)
#> An ingres object
#> --------------------
#> 16314 features across 90 samples
#> 9 clusters
#> Its Boolean network has 30 nodes
#> Viper not performed
```

Running VIPER

The next step is to run VIPER. For more information about the algorithm VIPER uses, please refer to its documentation.

Ingres runs VIPER on the provided gene expression matrix, using the regulon supplied by the user. This regulon can come from the ARACNe algorithm[4,5] or from other databases such as Dorothea[6]. If several regulons are provided, the metaVIPER algorithm [14] will be run instead.
This algorithm is designed to integrate multiple interactomes, not necessarily tissue-matched, as VIPER needed to work accurately. This is most valuable in single-cell RNA-seq data, due to its heterogeneity, inherent noisiness and low sequencing depth[7]. Indeed, metaVIPER has been shown to reduce bias and batch effect, generating reproducible protein-activity signatures [14]. Therefore, we recommend this method for Ingres. Collections of tumour regulons are available at Bioconductor[8]. The package `aracne.networks`, a suggested dependency for *Ingres*, contains ARACNe-inferred networks from TCGA tumor datasets.

VIPER produces a matrix of Normalised Enrichment Scores (NES) per each gene and cell.

```
# Using a single regulon for speed
ing = performViper(ing, aracne.networks::regulonblca)
ing
#> An ingres object
#> --------------------
#> 16314 features across 90 samples
#> 9 clusters
#> Its Boolean network has 30 nodes
#> Viper performed
```

Computing the PBNs

Now we can generate PBNs for every cell and/or every cluster in our *Ingres* object. This will generate a PBN that is a copy of the provided Boolean network but with a second rule in each gene node that is a fixed 1 or 0 (always active or always inactive). The probability of these rules being chosen will be proportional to the NES computed

by VIPER. By default, this step will re-scale NES to (-1,1) for each gene in each cell. This is, the gene with the highest NES in a given cell will be re-scaled to 1 and the gene with the lowest NES in that same cell will be re-scaled to -1, and every other gene will be re-scaled relative to that range. Then, all genes with a re-scaled NES>0 will have a fixed 1 rule added to the corresponding network node, with a probability equals to its re-scaled NES. All genes with a re-scaled NES<0 will have a fixed 0 ruled added, with a probability equals to [re-scaled NES]. The original rule for each node will have a probability equals to 1-[fixed rule probability].

The range (-1,1) can be changed by the user, in case a less intense effect by the RNA-seq data is desired. For example, if the range is inputed as (-0.5, 0.5), the gene with the highest expression will have a 0.5 probability of activating its fixed rule, the same chance as its original rule.

```
ing = computePbnBySingleCell(ing)
ing
#> An ingres object
#> --------------------
#> 16314 features across 90 samples
#> 9 clusters
#> Its Boolean network has 30 nodes
#> Viper performed
#> PBNs computed for each cell
head(ing@single.cell.pbn)
#>                          cell cluster         AKT        ATM      CHEK1_2
#> 1 sample1@AAGGAGCCAGCTTCGG-1       2  0.01817583 -0.7989694  0.099920471
#> 2 sample1@ACAGCTAAGATCCCGC-1       1 -0.07269741 -1.0000000  0.219362784
#> 3 sample1@ACGATGTAGTACATGA-1       3 -0.01603311 -0.9357372  0.513787703
#> 4 sample1@ACGGAGACAAATTGCC-1       2 -0.07899067 -0.7880775  0.055078000
#> 5 sample1@ACGGGCTCAAACCTAC-1       2 -0.04870924 -0.8023977 -0.002197616
#> 6 sample1@ACTGATGAGAGCTGGT-1       8  0.01102953 -0.6630933  0.153858311
#>     CyclinA    CyclinD1    CyclinE1        E2F1        EGFR       FGFR3
#> 1 1.0000000 0.005835620 -0.12590699  0.09456779 -0.002438333 -0.06437469
#> 2 0.2526539 0.100580602  0.03804537  0.14432723 -0.102653311 -0.08783756
#> 3 0.4896386 0.210391219 -0.04145495  1.00000000 -0.378800533  0.14514210
#> 4 1.0000000 0.060062911 -0.06046783 -0.06633384 -0.135412201 -0.03497538
#> 5 1.0000000 0.003268819  0.04396283  0.40529261  0.229916691 -0.24036193
#> 6 0.1336534 0.492846995 -0.26039627 -0.69295991 -0.509692556 -0.37852354
#>          GRB2         MDM2    p14ARF     p21CIP       PI3K        PTEN
#> 1  0.31113702 -0.033343239 0.3929228 0.38233341 -0.4071378  0.29342925
#> 2  0.32854332 -0.094203638 0.3691654 0.42156174 -0.7358486  0.35724316
#> 3 -0.07954781  0.054171975 0.5303340 0.01893101 -0.5368198 -0.05223321
#> 4  0.14676515 -0.033723994 0.4598230 0.29802852 -0.5897442  0.20687463
#> 5  0.05351330  0.008933491 0.3479199 0.37278246 -0.4302515  0.27558028
#> 6  0.37222843 -0.326267067 0.1103229 0.37679082 -1.0000000  0.05096307
#>         RAS         RB1        RBL2        SPRY        TP53
#> 1 0.3769146 -0.02103803 -0.2541966 -0.2188479 -0.10594942
#> 2 0.6074293 -0.15998474 -0.2668762 -0.2855518  0.10555768
#> 3 0.5145371 -0.38879500 -0.4278514 -0.3070425  0.15455175
#> 4 0.5643470 -0.18693366 -0.2616696 -0.1796798 -0.09579184
#> 5 0.5007909 -0.38687170 -0.3977755 -0.2124557  0.17860741
#> 6 0.4309240  0.66462958  0.3797944 -0.2995955 -0.12893251

head(computePbnBySingleCell(ing, c(-0.5, 0.5))@single.cell.pbn)
#>                          cell cluster          AKT        ATM      CHEK1_2
#> 1 sample1@AAGGAGCCAGCTTCGG-1       2  0.009087914 -0.3994847  0.049960235
#> 2 sample1@ACAGCTAAGATCCCGC-1       1 -0.036348706 -0.5000000  0.109681392
#> 3 sample1@ACGATGTAGTACATGA-1       3 -0.008016553 -0.4678686  0.256893851
#> 4 sample1@ACGGAGACAAATTGCC-1       2 -0.039495337 -0.3940388  0.027539000
#> 5 sample1@ACGGGCTCAAACCTAC-1       2 -0.024354619 -0.4011988 -0.001098808
#> 6 sample1@ACTGATGAGAGCTGGT-1       8  0.005514765 -0.3315466  0.076929156
#>     CyclinA    CyclinD1    CyclinE1        E2F1        EGFR       FGFR3
#> 1 0.5000000 0.002917810 -0.06295350  0.04728390 -0.001219166 -0.03218735
#> 2 0.1263270 0.050290301  0.01902269  0.07216362 -0.051326656 -0.04391878
#> 3 0.2448193 0.105195610 -0.02072748  0.50000000 -0.189400267  0.07257105
#> 4 0.5000000 0.030031455 -0.03023391 -0.03316692 -0.067706100 -0.01748769
#> 5 0.5000000 0.001634409  0.02198142  0.20264630  0.114958345 -0.12018097
#> 6 0.0668267 0.246423498 -0.13019813 -0.34647995 -0.254846278 -0.18926177
#>          GRB2         MDM2     p14ARF      p21CIP       PI3K         PTEN
#> 1  0.15556851 -0.016671619 0.19646142 0.191166703 -0.2035689  0.14671463
#> 2  0.16427166 -0.047101819 0.18458272 0.210780872 -0.3679243  0.17862158
#> 3 -0.03977391  0.027085987 0.26516699 0.009465505 -0.2684099 -0.02611660
#> 4  0.07338257 -0.016861997 0.22991148 0.149014262 -0.2948721  0.10343732
#> 5  0.02675665  0.004466745 0.17395994 0.186391231 -0.2151258  0.13779014
#> 6  0.18611422 -0.163133533 0.05516143 0.188395408 -0.5000000  0.02548154
#>         RAS         RB1        RBL2        SPRY        TP53
#> 1 0.1884573 -0.01051901 -0.1270983 -0.10942395 -0.05297471
#> 2 0.3037147 -0.07999237 -0.1334381 -0.14277588  0.05277884
#> 3 0.2572686 -0.19439750 -0.2139257 -0.15352125  0.07727588
#> 4 0.2821735 -0.09346683 -0.1308348 -0.08983989 -0.04789592
#> 5 0.2503954 -0.19343585 -0.1988878 -0.10622785  0.08930371
#> 6 0.2154620  0.33231479  0.1898972 -0.14979777 -0.06446626
```

We can also compute a single PBN per each cluster. In that case, the median NES for each gene across all cells in that cluster is used to produce the PBN:

```
ing = computePbnByCluster(ing)
ing
#> An ingres object
#> --------------------
#> 16314 features across 90 samples
#> 9 clusters
#> Its Boolean network has 30 nodes
#> Viper performed
#> PBNs computed for each cell
#> PBNs computed for each cluster
```

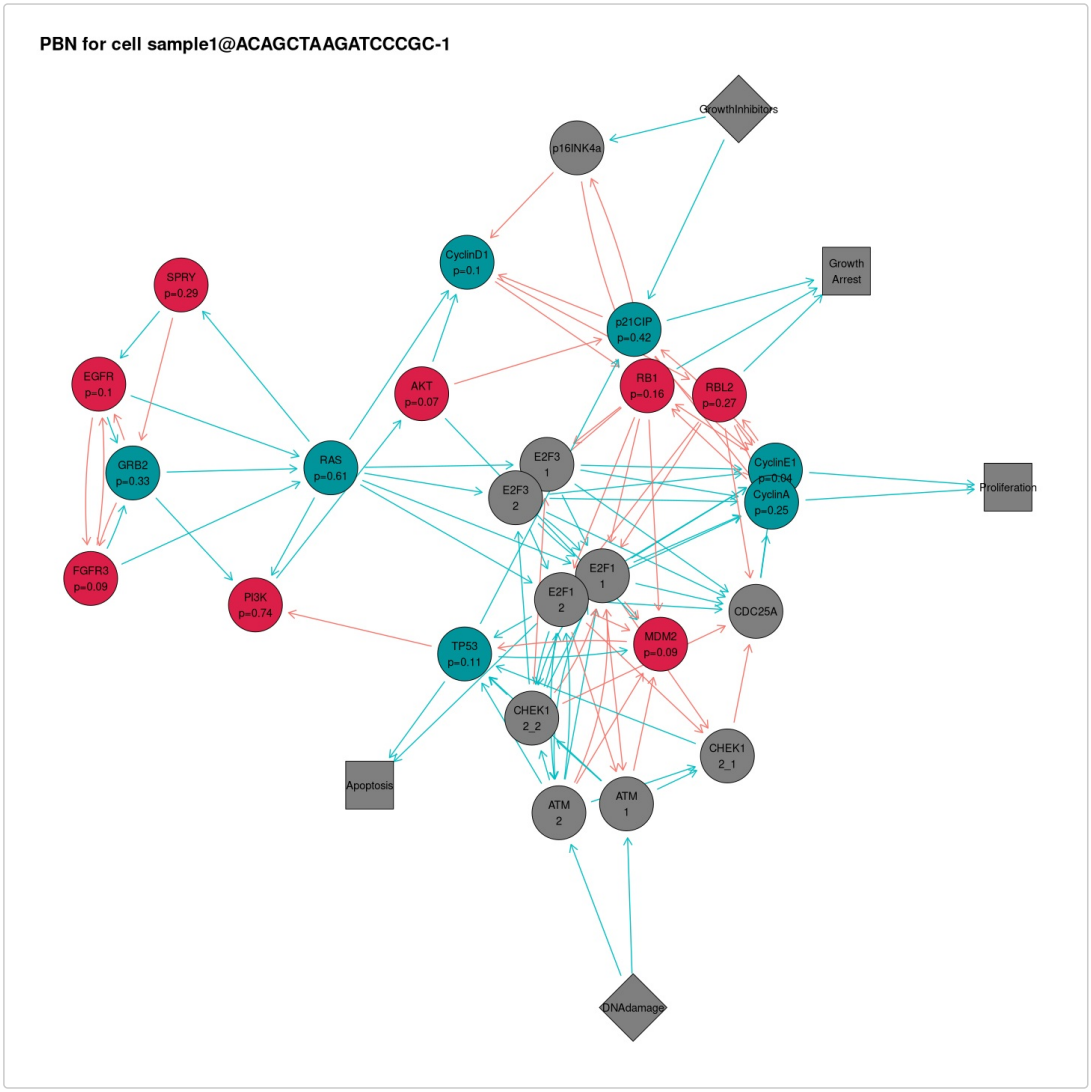
Visualization of results

We can export any PBN for further analysis by providing its cell barcode or cluster id:

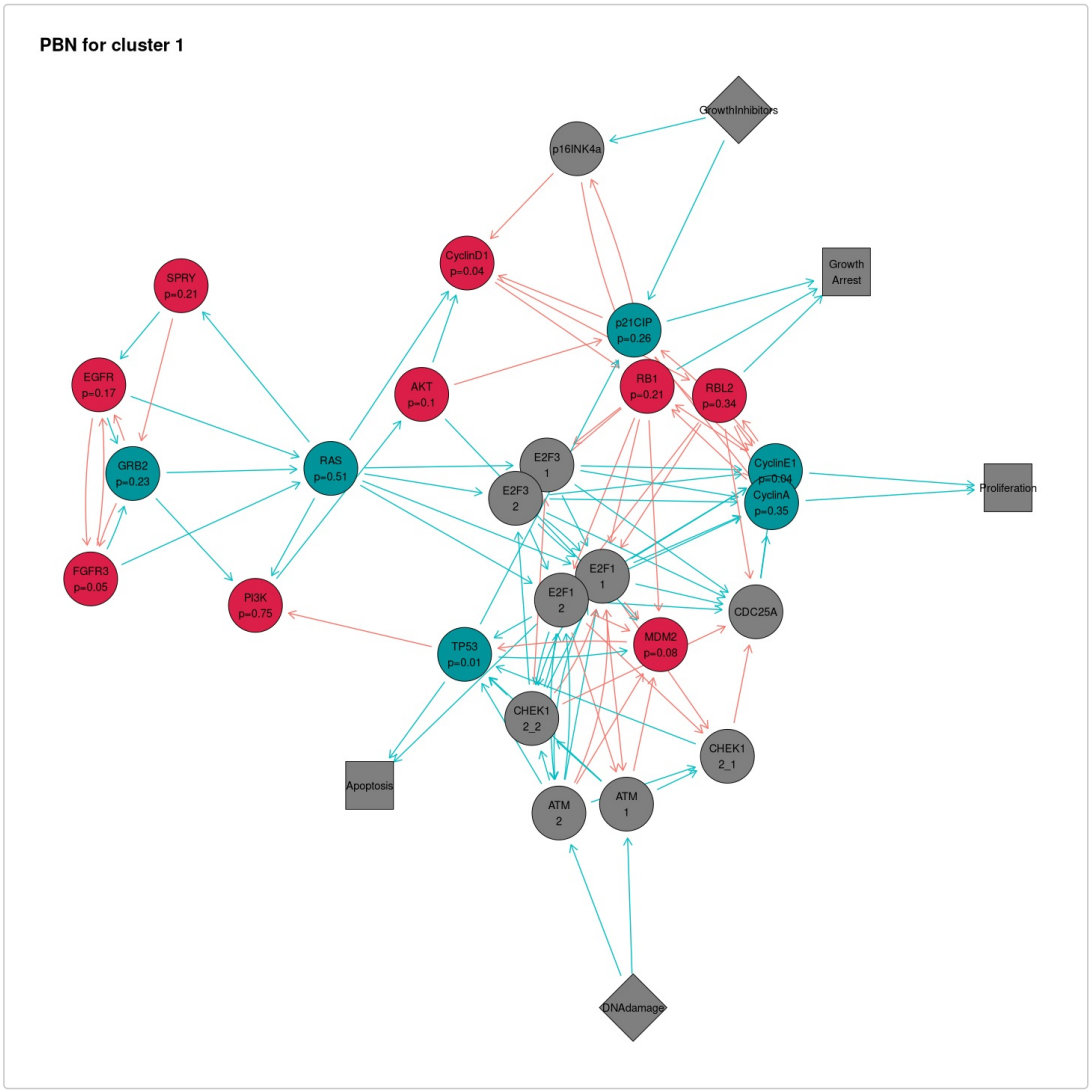
```
produceNetworkForCell(ing, "sample1@ACAGCTAAGATCCCGC-1")

produceNetworkForCluster(ing, "1")
```

In the same way, we can plot any network:

```
cellPbnPlot(ing, "sample1@ACAGCTAAGATCCCGC-1")
```

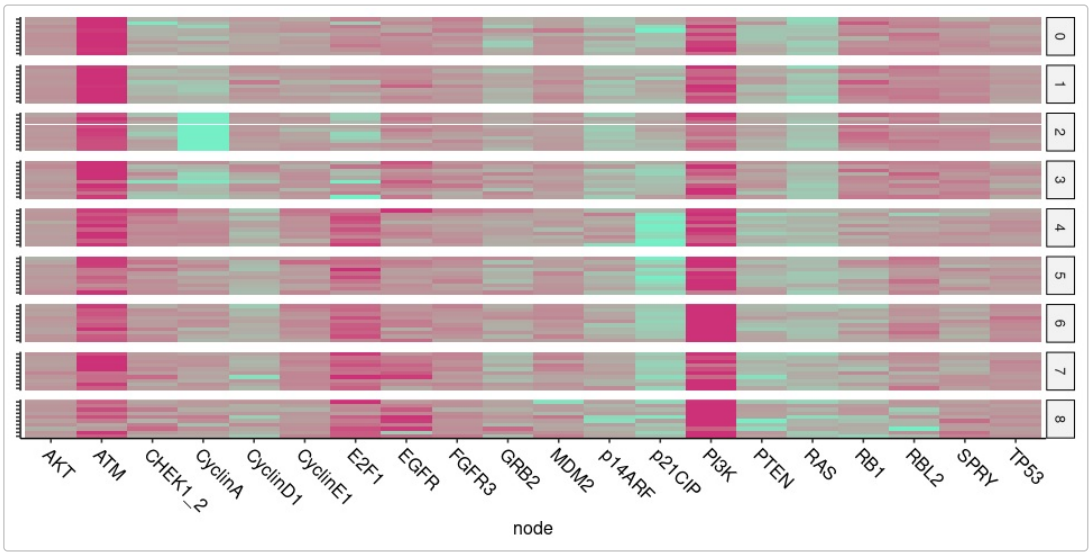



```
clusterPbnPlot(ing, "1")
```

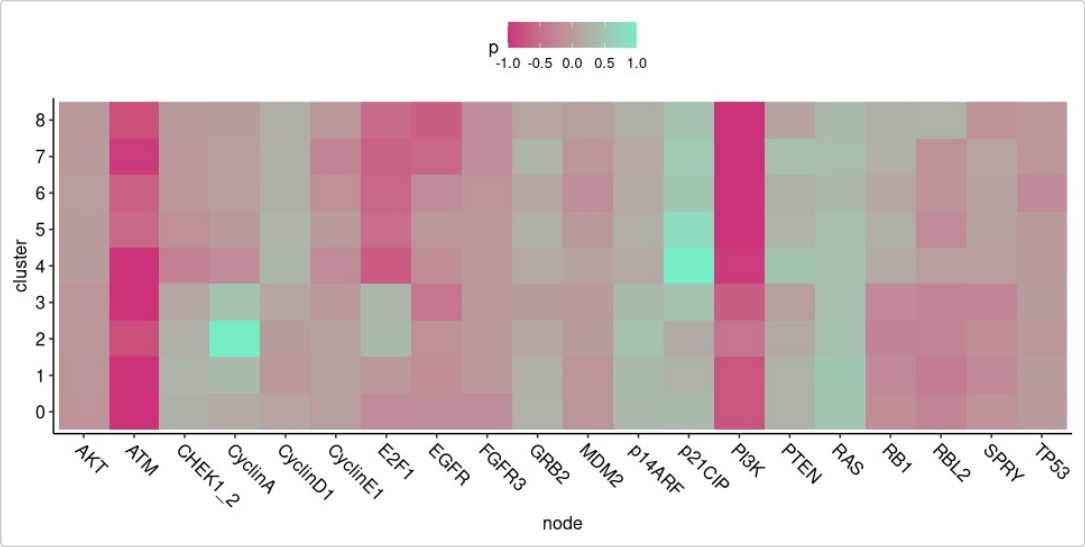


An overview of the results can be plotted as heatmaps:

```
cellGenesHeatmap(ing)
```




```
clusterGenesHeatmap(ing)
```



Finally, any network can be exported as a `BoolNet`[9] network. This package includes a variety of functions to analyse networks:

```
ing_network = produceNetworkForCluster(ing, "1")

boolnet_network = produceBoolnetNetwork(ing_network)

# Generate a initial state where all nodes have state 1
initial_state = rep(1, length(network))

# Compute one state transition from the given initial state.
transition1 = BoolNet::stateTransition(
  boolnet_network,
  type  = "probabilistic",
  state = initial_state)

head(transition1)
#>        DNAdamage GrowthInhibitors    Proliferation        Apoptosis
#>                1                1               1                1
#>    Growth_Arrest             EGFR
#>                1                0
```

SessionInfo

```
sessionInfo()
#> R version 4.2.1 (2022-06-23)
#> Platform: x86_64-pc-linux-gnu (64-bit)
#> Running under: Arch Linux
#>
#> Matrix products: default
#> BLAS:   /usr/lib/libblas.so.3.10.1
#> LAPACK: /usr/lib/liblapack.so.3.10.1
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
#>  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
#>  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
#>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
#> [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] ingres_1.0.0
#>
#> loaded via a namespace (and not attached):
#>   [1] utf8_1.2.2             reticulate_1.26        tidyselect_1.2.0
#>   [4] RSQLite_2.2.18         AnnotationDbi_1.58.0   htmlwidgets_1.5.4
#>   [7] grid_4.2.1             Rtsne_0.16             munsell_0.5.0
#>  [10] codetools_0.2-18       ica_1.0-3              future_1.28.0
#>  [13] miniUI_0.1.1.1         withr_2.5.0            spatstat.random_2.2-0
#>  [16] colorspace_2.0-3       progressr_0.11.0       Biobase_2.56.0
#>  [19] highr_0.9              knitr_1.40             rstudioapi_0.14
```

```
#>  [22] Seurat_4.2.0           stats4_4.2.1           ROCR_1.0-11
#>  [25] ggsignif_0.6.4         tensor_1.5             listenv_0.8.0
#>  [28] labeling_0.4.2         GenomeInfoDbData_1.2.8 polyclip_1.10-0
#>  [31] bit64_4.0.5            farver_2.1.1           parallelly_1.32.1
#>  [34] vctrs_0.4.2            generics_0.1.3         xfun_0.33
#>  [37] aracne.networks_1.22.0 R6_2.5.1               GenomeInfoDb_1.32.4
#>  [40] graphlayouts_0.8.2     bitops_1.0-7           spatstat.utils_2.3-1
#>  [43] cachem_1.0.6           assertthat_0.2.1       promises_1.2.0.1
#>  [46] scales_1.2.1           vroom_1.6.0            ggraph_2.1.0
#>  [49] rgeos_0.5-9            gtable_0.3.1           globals_0.16.1
#>  [52] goftest_1.2-3          tidygraph_1.2.2        rlang_1.0.6
#>  [55] splines_4.2.1          rstatix_0.7.0          lazyeval_0.2.2
#>  [58] spatstat.geom_2.4-0    broom_1.0.1            yaml_2.3.5
#>  [61] reshape2_1.4.4         abind_1.4-5            backports_1.4.1
#>  [64] httpuv_1.6.6           tools_4.2.1            ggplot2_3.3.6
#>  [67] ellipsis_0.3.2         spatstat.core_2.4-4    jquerylib_0.1.4
#>  [70] RColorBrewer_1.1-3     proxy_0.4-27           BiocGenerics_0.42.0
#>  [73] ggridges_0.5.4         Rcpp_1.0.9             plyr_1.8.7
#>  [76] zlibbioc_1.42.0        purrr_0.3.5            RCurl_1.98-1.9
#>  [79] ggpubr_0.4.0           rpart_4.1.16           deldir_1.0-6
#>  [82] pbapply_1.5-0          viridis_0.6.2          cowplot_1.1.1
#>  [85] S4Vectors_0.34.0       zoo_1.8-11             SeuratObject_4.1.2
#>  [88] ggrepel_0.9.1          cluster_2.1.3          magrittr_2.0.3
#>  [91] data.table_1.14.2      scattermore_0.8        lmtest_0.9-40
#>  [94] RANN_2.6.1             fitdistrplus_1.1-8     matrixStats_0.62.0
#>  [97] hms_1.1.2              patchwork_1.1.2        mime_0.12
#> [100] evaluate_0.17          xtable_1.8-4           XML_3.99-0.11
#> [103] IRanges_2.30.1         gridExtra_2.3          compiler_4.2.1
#> [106] tibble_3.1.8           KernSmooth_2.23-20     crayon_1.5.1
#> [109] htmltools_0.5.3        segmented_1.6-0        mgcv_1.8-40
#> [112] later_1.3.0            tzdb_0.3.0             tidyr_1.2.1
#> [115] DBI_1.1.3              tweenr_2.0.2           MASS_7.3-57
#> [118] Matrix_1.5-1           car_3.1-0              readr_2.1.3
#> [121] cli_3.4.1              parallel_4.2.1         igraph_1.3.5
#> [124] pkgconfig_2.0.3        sp_1.5-0               plotly_4.10.0.9001
#> [127] spatstat.sparse_2.1-1  bslib_0.4.0            XVector_0.36.0
#> [130] viper_1.30.0           stringr_1.4.1          digest_0.6.29
#> [133] sctransform_0.3.5      RcppAnnoy_0.0.19       spatstat.data_2.2-0
#> [136] Biostrings_2.64.1      BoolNet_2.1.7          rmarkdown_2.17
#> [139] leiden_0.4.3           uwot_0.1.14            kernlab_0.9-31
#> [142] shiny_1.7.2            lifecycle_1.0.3        nlme_3.1-157
#> [145] jsonlite_1.8.0         carData_3.0-5          viridisLite_0.4.1
#> [148] fansi_1.0.3            pillar_1.8.1           lattice_0.20-45
#> [151] KEGGREST_1.36.3        fastmap_1.1.0          httr_1.4.4
#> [154] survival_3.3-1         glue_1.6.2             conflicted_1.1.0
#> [157] png_0.1-7              bit_4.0.4              ggforce_0.4.1
#> [160] class_7.3-20           stringi_1.7.8          sass_0.4.2
#> [163] mixtools_1.2.0         blob_1.2.3             org.Hs.eg.db_3.15.0
#> [166] memoise_2.0.1          dplyr_1.0.10           irlba_2.3.5.1
#> [169] e1071_1.7-11           future.apply_1.9.1
```

References

1. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48**, 838–847 (2016).
2. Victori, P. & Buffa, F. M. Ingres: From single-cell RNA-seq data to single-cell probabilistic Boolean networks. 2022.09.04.506528 (2022) doi:10.1101/2022.09.04.506528.
3. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. (2021) doi:10.1016/j.cell.2021.04.048.
4. Basso, K. *et al.* Reverse engineering of regulatory networks in human B cells. *Nature Genetics* **37**, 382–390 (2005).
5. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).
6. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research* **29**, 1363–1375 (2019).
7. Ding, H. *et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nature Communications* **9**, 1471 (2018).
8. Giorgi, F. M. *Aracne.networks: ARACNe-inferred gene networks from TCGA tumor datasets*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.ARACNE.NETWORKS.
9. Müssel, C., Hopfensitz, M. & Kestler, H. A. BoolNet – an r package for generation, reconstruction and analysis of boolean networks. *Bioinformatics* **26**, 1378–1380 (2010).

Package 'ingres'

October 19, 2022

Title Infer Gene Probabilistic Boolean Networks from Single-Cell Data

Version 1.0.0

Description Given a gene regulatory boolean network and a RNA-seq dataset,
this package computes protein activity normalised enrichment scores
using 'VIPER', and then produces a probabilistic network using the scores
as probabilities for fixed node activation or deactivation,
in addition to the original Boolean functions.
For more information, refer to the preprint:
Victori and Buffa (2022) <doi:10.1101/2022.09.04.506528>.

License GPL-3

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

biocViews NetworkInference, GraphAndNetwork, NetworkEnrichment, Network,
GeneRegulation, GeneExpression, SingleCell, SystemsBiology, Software

Depends R (>= 4.2.0)

Imports methods,
viper,
tidygraph,
AnnotationDbi,
org.Hs.eg.db,
dplyr,
tidyr,
tidyselect,
scales,
ggplot2,
ggraph,
stringr,
ggpubr,
tibble,
magrittr,
igraph,
purrr,

readr,
usethis,
rlang,
Seurat

Suggests testthat (>= 3.0.0),
shiny,
plotly (>= 4.10.0),
aracne.networks,
BoolNet,
covr,
withr,
knitr,
rmarkdown

Config/testthat/edition 3

URL https://github.com/CBigOxf/ingres,
https://cbigoxf.github.io/ingres/

BugReports https://github.com/CBigOxf/ingres/issues

VignetteBuilder knitr

Language en-GB

R topics documented:

cellGenesHeatmap . . . 3
cellPbnPlot . . . 4
clusterGenesHeatmap . . . 4
clusterPbnPlot . . . 5
computePbnByCluster . . . 6
createExampleIngresObject . . . 7
createIngresObject . . . 7
createIngresObjectFromSeurat . . . 8
createNetworkGenesTemplate . . . 9
getClusterPbns . . . 10
getExpressionMatrix . . . 11
getIdents . . . 11
getNetwork . . . 12
getNetworkGenes . . . 12
getSingleCellPbns . . . 13
getViperResults . . . 13
ginmlToGraphml . . . 14
graphmlAsTidy . . . 15
ingres-class . . . 15
network . . . 16
network_genes . . . 16
performViper . . . 17
plotSelectedCell . . . 17

printAllNodes 18
produceBoolnetNetwork 19
produceNetworkForCell 19
produceNetworkForCluster 20
small_blca_wang 21
viper_results 21

Index **22**

`cellGenesHeatmap` *Plot a cell heatmap*

Description

Plots the probabilities for the identity function produced by `computePbnBySingleCell` as a heatmap, with nodes on the x axis and cells on the y axis, grouped by cluster, and returns it as a ggplot object.

Usage

```
cellGenesHeatmap(ingres.object)
```

Arguments

`ingres.object`
: An `ingres` object with a non-null `single.cell.pbn` slot.

Value

A `ggplot` object containing the cell PBN heatmap plot.

Examples

```
# Create an ingres object with viper slot
ing <- createExampleIngresObjectWithViper()
ing <- computePbnBySingleCell(ing)
cellGenesHeatmap(ing)
```

cellPbnPlot *Plot a cell PBN*

Description

Plots the PBN produced by computePbnBySingleCell for a given cell and returns it as a ggplot object

Usage

```
cellPbnPlot(ingres.object, cell.id)
```

Arguments

ingres.object	An ingres object with a non-null single.cell.pbn slot.
cell.id	A string containing the id of the cell to be plotted.

Value

A ggplot object containing the cell PBN plot.

Examples

```
createExampleIngresObjectWithViper
ing <- computePbnBySingleCell(ing)
cellPbnPlot(ing, "sample1@ACAGCTAAGATCCCGC-1")
```

clusterGenesHeatmap *Plot a cluster heatmap*

Description

Plots the probabilities for the identity function produced by computePbnByCluster as a heatmap, with nodes on the x axis and clusters on the y axis, and returns it as a ggplot object.

Usage

```
clusterGenesHeatmap(ingres.object)
```

Arguments

ingres.object	An ingres object with a non-null cluster.pbn slot.

Value

A ggplot object containing the cluster PBN heatmap plot.

Examples

```
ing <- createExampleIngresObjectWithViper()
ing <- computePbnByCluster(ing)
clusterGenesHeatmap(ing)
```

---

clusterPbnPlot *Plot a cluster PBN*

Description

Plots the PBN produced by computePbnByCluster for a given cluster and returns it as a ggplot object

Usage

```
clusterPbnPlot(ingres.object, cluster.id)
```

Arguments

ingres.object	An ingres object with a non-null cluster.pbn slot.
cluster.id	A string containing the id of the cluster to be plotted.

Value

A ggplot object containing the cluster PBN plot.

Examples

```
createExampleIngresObjectWithViper
ing <- computePbnByCluster(ing)
clusterPbnPlot(ing, "1")
```

computePbnByCluster *Compute a Probabilistic Boolean Network (PBN) for each cluster or each cell*

Description

After running performViper, this can be run to produce a PBN with the identity function probability for each node relative to the VIPER normalised enrichment score (NES) for cell. For computePbnByCluster, the median NES for all cells in that cluster will be used.

Usage

```
computePbnByCluster(ingres.object, range = c(-1, 1))

computePbnBySingleCell(ingres.object, range = c(-1, 1))
```

Arguments

ingres.object	An ingres object with a non-null VIPER slot.
range	A vector representing the range to which the NES are to be re-scaled. range[1]+range[2] == 0 should be TRUE. Defaults to [-1, 1]

Value

An ingres object with the cluster.pbn or the single.cell.pbn slot filled

Functions

- computePbnByCluster: Compute a Probabilistic Boolean Network (PBN) for each cluster
- computePbnBySingleCell: Compute a Probabilistic Boolean Network (PBN) for each cell

Examples

```
# Create an ingres object with viper slot
ing <- createExampleIngresObjectWithViper()

ing <- computePbnByCluster(ing)
head(ing@cluster.pbn)

ing <- computePbnBySingleCell(ing)
head(ing@single.cell.pbn)

# Restrict range to (-0.5, 0.5)
ing <- computePbnByCluster(ing, range = c(-0.5, 0.5))
```

```
head(ing@cluster.pbn)

ing <- computePbnBySingleCell(ing, range = c(-0.5, 0.5))
head(ing@single.cell.pbn)
```

`createExampleIngresObject` *Create a toy* `ingres` *object to use in examples and tests.*

Description

`createExampleIngresObjectWithViper` additionally fills the `viper` slot with a dummy `viper` object

Usage

```
createExampleIngresObject()

createExampleIngresObjectWithViper()
```

Value

An `ingres` object

Examples

```
createExampleIngresObject()
createExampleIngresObjectWithViper()
```

`createIngresObject` *Create a new* `ingres` *object.*

Description

Create a new `ingres` object.

Usage

```
createIngresObject(expression.matrix, idents, network.genes, network)
```

Arguments

`expression.matrix`	A single-cell expression matrix, with cells in rows and genes in columns.
`idents`	A data frame with a column for cell barcode and another column for the corresponding cluster or subpopulation.
`network.genes`	A data frame with a column for node names and another column for the corresponding entrez IDs.
`network`	A object of class `tbl_graph` representing our network of interest.

Value

An `ingres` object.

Examples

```
# Get expression matrix from small_blca_wang for convenience, but it can be
# any single-cell expression matrix, from any source. Same for idents.
expression.matrix <-
  as.matrix(Seurat::GetAssayData(
    small_blca_wang,
    assay = "RNA", slot = "data"
  ))

idents <- data.frame(
  cell = names(Seurat::Idents(small_blca_wang)),
  cluster = as.character(Seurat::Idents(small_blca_wang)),
  check.names = FALSE
)

ing <- createIngresObject(expression.matrix, idents, network_genes, network)
ing
```

`createIngresObjectFromSeurat` *Create a new* `ingres` *object using the expression data in a* `Seurat` *object.*

Description

Create a new `ingres` object using the expression data in a `Seurat` object.

Usage

```
createIngresObjectFromSeurat(
  seurat.object,
  seurat.assay = "RNA",
  slot = "data",
  network.genes,
  network
)
```

Arguments

`seurat.object`	A `Seurat` object with QC, scaling and clustering already executed.
`seurat.assay`	The name of the assay to be imported, normally 'RNA' or 'Spatial'.
`slot`	The name of the slot that contains the matrix to be used.
`network.genes`	A data frame with a column for node names and another column for the corresponding entrez IDs.
`network`	A object of class `tbl_graph` representing our network of interest.

Value

An `ingres` object.

Examples

```
createIngresObjectFromSeurat(
  small_blca_wang, "RNA", "data", network_genes, network
)
```

`createNetworkGenesTemplate`

Create a network genes data frame. Optionally store it as csv and open it for editing To create an ingres object, a data frame with the network nodes and the corresponding gene symbols must be provided. This function simplifies the process. If the gene nodes are correct gene symbols, then modification is not needed and the returned data frame can be directly passed to the ingres constructors.

Description

Create a network genes data frame. Optionally store it as csv and open it for editing To create an ingres object, a data frame with the network nodes and the corresponding gene symbols must be provided. This function simplifies the process. If the gene nodes are correct gene symbols, then modification is not needed and the returned data frame can be directly passed to the ingres constructors.

Usage

```
createNetworkGenesTemplate(network, dir = getwd(), store = TRUE, modify = TRUE)
```

Arguments

network	A tidygraph network.
dir	The directory where the csv will be stored, if applicable.
store	If true, store the data frame as a csv
modify	If true, and store is also true, open it to be modified by the user

Value

The template data frame.

Examples

```
createNetworkGenesTemplate(network, store = FALSE)
```

---

getClusterPbns *Get the cluster PBNs from an* ingres *object*

Description

Get the cluster PBNs from an ingres object

Usage

```
getClusterPbns(ingres.object)
```

Arguments

ingres.object	an ingres object

Value

A data frame containing the generated cluster PBNs

Examples

```
ing <- createExampleIngresObjectWithViper()
ing <- ing %>% computePbnByCluster()
head(getClusterPbns(ing))
```

getExpressionMatrix *Get the Expression Matrix from an* ingres *object*

Description

Get the Expression Matrix from an ingres object

Usage

```
getExpressionMatrix(ingres.object)
```

Arguments

ingres.object	an ingres object

Value

An expression matrix

Examples

```
ing <- createExampleIngresObject()
head(GetExpressionMatrix(ing))
```

getIdents *Get the idents from an* ingres *object*

Description

Get the idents from an ingres object

Usage

```
getIdents(ingres.object)
```

Arguments

ingres.object	an ingres object

Value

A data frame with a column for cell barcode and another column for the corresponding cluster or subpopulation.

Examples

```
ing <- createExampleIngresObject()
head(GetIdents(ing))
```

`getNetwork` *Get the network from an* `ingres` *object*

Description

Get the network from an `ingres` object

Usage

```
getNetwork(ingres.object)
```

Arguments

`ingres.object`	an `ingres` object

Value

The network as a object of class `tbl_graph`

Examples

```
ing <- createExampleIngresObject()
head(getNetwork(ing))
```

`getNetworkGenes` *Get the network genes from an* `ingres` *object*

Description

Get the network genes from an `ingres` object

Usage

```
getNetworkGenes(ingres.object)
```

Arguments

`ingres.object`	an `ingres` object

Value

A data frame with a column for node names and another column for the corresponding entrez IDs.

Examples

```
ing <- createExampleIngresObject()
head(tetNetworkGenes(ing))
```

---

getSingleCellPbns *Get the single-cell PBNs from an* ingres *object*

Description

Get the single-cell PBNs from an ingres object

Usage

```
getSingleCellPbns(ingres.object)
```

Arguments

ingres.object	an ingres object

Value

A data frame containing the generated single-cell PBNs

Examples

```
ing <- createExampleIngresObjectWithViper()
ing <- ing %>% computePbnBySingleCell()
head(getSingleCellPbns(ing))
```

---

getViperResults *Get the viper results from an* ingres *object*

Description

Get the viper results from an ingres object

Usage

```
getViperResults(ingres.object)
```

Arguments

ingres.object	an ingres object

Value

A data frame containing viper results

Examples

```
ing <- createExampleIngresObjectWithViper()
head(getViperResults(ing))
```

---

ginmlToGraphml

Convert a GinSim file into a GraphML file GinSim files have the extension .zginml. This utility function converts such files into the GraphML format. keeping the kind - fate, input or gene -, the edge sign and the rule - formulae - data.

Description

Convert a GinSim file into a GraphML file GinSim files have the extension .zginml. This utility function converts such files into the GraphML format. keeping the kind - fate, input or gene -, the edge sign and the rule - formulae - data.

Usage

```
ginmlToGraphml(ginzipFile, fates = c(), dest = NULL)
```

Arguments

ginzipFile	The path to the .zginml file. The GraphML file will be created in the same directory.
fates	A vector of fate names. If left empty, both fate and gene nodes will be labelled as kind='gene'
dest	The path to the graphml file that will be created. If NULL, defaults to the same path as the zginml file, but with the graphml extension.

Value

A vector with the lines of the newly created GraphML file.

Examples

```
filename <-
  system.file("extdata", "example_ginsim.zginml", package = "ingres")
head(ginmlToGraphml(filename))
```

graphmlAsTidy *Convert GraphML format to tidygraph*

Description

In order to be used by ingres, GraphML files have to be converted to tidygraph format. This helper does that.

Usage

```
graphmlAsTidy(filename)
```

Arguments

filename	The path to the the GraphML file.

Value

A tidygraph object

Examples

```
filename <-
  system.file("extdata", "example_network.graphml", package = "ingres")
graphmlAsTidy(filename)
```

ingres-class *The ingres class.*

Description

The ingres class.

Value

An `ingres` object.

Slots

`expression` matrix.
`idents` data.frame.
`network.genes` data.frame.
`viper` data.frame.
`cluster.pbn` data.frame.
`single.cell.pbn` data.frame.
`network` `tbl_graph`.

`network` *An example network, in the format required by the package. The network is a modified version of the one in Remy et al. Cancer Res 75, 4042–4052 (2015)*

Description

An example network, in the format required by the package. The network is a modified version of the one in Remy et al. Cancer Res 75, 4042–4052 (2015)

Usage

```
network
```

Format

An object of class `tbl_graph` (inherits from `igraph`) of length 30.

`network_genes` *An example dataframe with the genes contained in the network, in the format required by the package*

Description

An example dataframe with the genes contained in the network, in the format required by the package

Usage

```
network_genes
```

Format

An object of class `data.frame` with 28 rows and 2 columns.

`performViper` *Perform VIPER*

Description

Performs VIPER analysis on a RNA-seq dataset contained by an ingres object. This function is merely a wrapper around the `original VIPER function`, designed to suit the ingres workflow and properly prepare the input data.

Usage

```
performViper(ingres.object, regulon, verbose = FALSE)
```

Arguments

`ingres.object`	An ingres object.
`regulon`	Object of class regulon or list of objects of class regulon for metaVIPER analysis
`verbose`	Logical, whether progression messages should be printed in the terminal

Value

An ingres object with the result of the VIPER inference in the `viper` slot

Examples

```
# Subset regulon to speed up example run
regulon <- aracne.networks::regulonblca[1:100]
ing <- createExampleIngresObject()
performViper(ing, regulon)
```

`plotSelectedCell` *Plot the PBN of a cell selected in an interactive plot.*

Description

Launches a Shiny app with an interactive tSNE plot from a Seurat object. When a cell is clicked, the PBN for that cell is plotted side by side with the tSNE plot. Internally, it uses `cellPbnPlot` on the selected cell.

Usage

```
plotSelectedCell(ingres.object, seurat.object)
```

Arguments

ingres.object	An ingres object with a non-null single.cell.pbn slot.
seurat.object	A Seurat object created with the same dataset as the ingres object. Ideally, this object was used to create the ingres object through createIngresObjectFromSeurat.

Value

No return value, called only to launch the Shiny app.

Examples

```
ing <- createIngresObjectFromSeurat(
  small_blca_wang, "RNA", "data", network_genes, network
)
ing@viper <- viper_results
ing <- computePbnBySingleCell(ing)
plotSelectedCell(ing, small_blca_wang)
```

---

printAllNodes *Print all nodes in a network For testing and checks purposes.*

Description

Print all nodes in a network For testing and checks purposes.

Usage

```
printAllNodes(network)
```

Arguments

network	The network which nodes will be printed.

Value

No return value, called only to print into the output.

Examples

```
printAllNodes(network)
```

`produceBoolnetNetwork` *Convert an ingres network into a BoolNet one*

Description

Convert an ingres network into a BoolNet one

Usage

```
produceBoolnetNetwork(network)
```

Arguments

`network`	The network to be converted, created by `produceNetworkForCluster()` or `produceNetworkForCell()`

Value

A BoolNet object.

Examples

```
# Create an ingres object with viper slot
ing <- createIngresObjectFromSeurat(
  small_blca_wang, "RNA", "data", network_genes, network
)
ing@viper <- viper_results

# Compute PBNs by cluster
ing <- computePbnByCluster(ing)

# Produce a network for an arbitrary cluster
network <- produceNetworkForCluster(ing, "1")

produceBoolnetNetwork(network)
```

`produceNetworkForCell` *Produce a tidygraph network for a given cell*

Description

Produce a tidygraph network for a given cell

Usage

```
produceNetworkForCell(ingres.object, cell.id)
```

Arguments

ingres.object	The ingres object containing the relevant data.
cell.id	The id of the cell for which the network is to be produced.

Value

A tidygraph object containing the network for the given cell.

Examples

```
# Create an ingres object with viper slot
ing <- createIngresObjectFromSeurat(
  small_blca_wang, "RNA", "data", network_genes, network
)
ing@viper <- viper_results

ing <- computePbnBySingleCell(ing)

network <- produceNetworkForCell(ing, "sample1@ACAGCTAAGATCCCGC-1")
print(network)
```

---

produceNetworkForCluster

Produce a tidygraph network for a given cluster

Description

Produce a tidygraph network for a given cluster

Usage

```
produceNetworkForCluster(ingres.object, cluster.id)
```

Arguments

ingres.object	The ingres object containing the relevant data.
cluster.id	The id of the cluster for which the network is to be produced.

Value

A tidygraph object containing the network for the given cluster.

Examples

```
# Create an ingres object with viper slot
ing <- createIngresObjectFromSeurat(
  small_blca_wang, "RNA", "data", network_genes, network
)
ing@viper <- viper_results

ing <- computePbnByCluster(ing)

network <- produceNetworkForCluster(ing, "1")
print(network)
```

`small_blca_wang` *An example small Seurat object, using dataset GSE130001 from Wang et al. Genome Med. 2020 Dec;12(1):24.*

Description

An example small Seurat object, using dataset GSE130001 from Wang et al. Genome Med. 2020 Dec;12(1):24.

Usage

```
small_blca_wang
```

Format

An object of class `Seurat` with 16314 rows and 90 columns.

`viper_results` *The results of running* `performViper()` *on an ingres object created with* `small_blca_wang`*, intended to speed up vignettes and examples. This data frame was subset to only contain the genes in the* `network_genes` *file, so to keep the file size small.*

Description

The results of running `performViper()` on an ingres object created with `small_blca_wang`, intended to speed up vignettes and examples. This data frame was subset to only contain the genes in the `network_genes` file, so to keep the file size small.

Usage

```
viper_results
```

Format

An object of class `data.frame` with 90 rows and 25 columns.

Index

∗ **datasets**
network, 16
network_genes, 16
small_blca_wang, 21
viper_results, 21

cellGenesHeatmap, 3
cellPbnPlot, 4, *17*
clusterGenesHeatmap, 4
clusterPbnPlot, 5
computePbnByCluster, *4, 5*, 6
computePbnBySingleCell, *3, 4*
computePbnBySingleCell *(computePbnByCluster)*, 6
createExampleIngresObject, 7
createExampleIngresObjectWithViper *(createExampleIngresObject)*, 7
createIngresObject, 7
createIngresObjectFromSeurat, 8, *18*
createNetworkGenesTemplate, 9

getClusterPbns, 10
getExpressionMatrix, 11
getIdents, 11
getNetwork, 12
getNetworkGenes, 12
getSingleCellPbns, 13
getViperResults, 13
ginmlToGraphml, 14
graphmlAsTidy, 15

ingres, *3–7*, *10–14*, *18*
ingres *(ingres-class)*, 15
ingres-class, 15
ingresClass *(ingres-class)*, 15
ingresObject *(ingres-class)*, 15

network, 16
network_genes, 16

performViper, *6*, 17
performViper(), *21*
plotSelectedCell, 17
printAllNodes, 18
produceBoolnetNetwork, 19
produceNetworkForCell, 19
produceNetworkForCell(), *19*
produceNetworkForCluster, 20
produceNetworkForCluster(), *19*

small_blca_wang, 21

viper_results, 21

B

Settings for all *microCJ* figures in Chapter 4

B.1 Default settings

Unless otherwise stated, these settings were used for all simulations:

Table B.1: General settings

Setting	**Value**
replicates	100
time steps	3000
starting-cells	200
cell-radius	10 µm
graph-file	network file containing the modified MAPK network, as showed in Figure 3.3.
decision-window	200
arrest-time	5
instant-arrest	false
initial-states	random
diffusion-grid-size	4
diffusion-voxel-size	20 µm
shoving	false
random-mutations	0
close-group-placement	false
diffusion	true
distributor-2d	hexagonal
distributor-3d	random
cache-positions	false
network	sync

Table B.2: Diffusion settings

Substance	**Setting**	**Value**
Oxygen	name	oxygen
Oxygen	type	nutrient
Oxygen	input-node	oxygen_supply
Oxygen	initial-concentration	0.28 mM
Oxygen	diffusion-coefficient	10^{-9} m^2/s
Oxygen	threshold	0.02 mM
Oxygen	consumption	0.005 mol/s/cell
Oxygen	production	
EGF	name	EGF
EGF	type	factor
EGF	input-node	EGFR_stimulus
EGF	initial-concentration	0 mM
EGF	diffusion-coefficient	$2.7 * 10^{-11}$ m^2/s
EGF	threshold	0.0004 mM
EGF	consumption	0.0004 mol/s/cell.
EGF	production	0.0004 mol/s/cell.

B.2 Specific settings

For the specific settings, only the parameters with different values to the ones in the default settings are mentioned.

Table B.3: Settings by figure

Figure	Setting	Value
Figure 4.3 and related	starting-cells	100
Figure 4.9 and related (hypoxia)	Oxygen initial-concentration	0.04 mM
Figure 4.15 and related (no EGF production)	EGF production	0 m^2/s
Figure 4.17 and related (no EGF production)	EGF production	0 m^2/s
Figure 4.17 and related (hypoxia)	Oxygen initial-concentration	0.04 mM
Figure 4.19 and related	random-mutations	0.8
Figure 4.19 and related (DNA damage)	DNA_damage fixed_input	true
Figure 4.22 and related	random-mutations	0.8
Figure 4.22 and related	distributor-2d	hexagonal

Table B.4: Diffusion settings for drugs (figure 4.25 and related)

Setting	Value
type	drug
input-node	Varies
initial-concentration	0 mM
dose-concentration	1E-2 mM
diffusion-coefficient	$2.5 * 10^{-10}$ m^2/s
threshold	$2 * 10^{-5}$ mM
consumption	$4 * 10^{-7}$ mol/s/cell
start-time	200
end-time	2000

References

1. Savage, P. *et al.* A Targetable EGFR-Dependent Tumor-Initiating Program in Breast Cancer. *Cell Reports* **21,** 1140–1149. doi:10.1016/j.celrep.2017.10.015 (Oct. 2017).
2. Victori, P. & Buffa, F. M. The Many Faces of Mathematical Modelling in Oncology. *The British Journal of Radiology,* 20180856. doi:10.1259/bjr.20180856 (Nov. 2018).
3. Victori, P. & Buffa, F. M. *Ingres: From Single-Cell RNA-seq Data to Single-Cell Probabilistic Boolean Networks* Sept. 2022. doi:10.1101/2022.09.04.506528.
4. Bray, F. *et al.* Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **68,** 394–424. doi:10.3322/caac.21492 (2018).
5. Collignon, J., Lousberg, L., Schroeder, H. & Jerusalem, G. Triple-Negative Breast Cancer: Treatment Challenges and Solutions. *Breast Cancer: Targets and Therapy,* 93. doi:10.2147/BCTT.S69488 (May 2016).
6. Lim, E., Metzger-Filho, O. & Winer, E. P. The Natural History of Hormone Receptor-Positive Breast Cancer. *Oncology* **26,** 688–94, 696 (Aug. 2012).
7. Lin, N. U. *et al.* Sites of Distant Recurrence and Clinical Outcomes in Patients with Metastatic Triple-Negative Breast Cancer. *Cancer* **113,** 2638–2645. doi:10.1002/cncr.23930 (2008).
8. Yin, L., Duan, J.-J., Bian, X.-W. & Yu, S.-c. Triple-Negative Breast Cancer Molecular Subtyping and Treatment Progress. *Breast Cancer Research* **22,** 61. doi:10.1186/s13058-020-01296-5 (Dec. 2020).
9. Dent, R. *et al.* Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. *Clinical Cancer Research* **13,** 4429–4434. doi:10.1158/1078-0432.CCR-06-3045 (Aug. 2007).
10. da Silva, J. L., Cardoso Nunes, N. C., Izetti, P., de Mesquita, G. G. & de Melo, A. C. Triple Negative Breast Cancer: A Thorough Review of Biomarkers. *Critical Reviews in Oncology/Hematology* **145,** 102855. doi:10.1016/j.critrevonc.2019.102855 (Jan. 2020).
11. Howlader, N. *et al.* US Incidence of Breast Cancer Subtypes Defined by Joint Hormone Receptor and HER2 Status. *JNCI: Journal of the National Cancer Institute* **106,** dju055. doi:10.1093/jnci/dju055 (May 2014).
12. Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A. & Caggiano, V. Descriptive Analysis of Estrogen Receptor (ER)-Negative, Progesterone Receptor (PR)-Negative, and HER2-negative Invasive Breast Cancer, the so-Called Triple-Negative Phenotype. *Cancer* **109,** 1721–1728. doi:10.1002/cncr.22618 (2007).
13. Hines, L. M. *et al.* Ethnic Disparities in Breast Tumor Phenotypic Subtypes in Hispanic and Non-Hispanic White Women. *Journal of Women's Health* **20,** 1543–1550. doi:10.1089/jwh.2010.2558 (Oct. 2011).
14. Ding, Y. C. *et al.* Molecular Subtypes of Triple-Negative Breast Cancer in Women of Different Race and Ethnicity. *Oncotarget* **10,** 198–208. doi:10.18632/oncotarget.26559 (Jan. 2019).
15. Lyons, T. G. Targeted Therapies for Triple-Negative Breast Cancer. *Current Treatment Options in Oncology* **20,** 82. doi:10.1007/s11864-019-0682-x (Nov. 2019).
16. Sledge, G. W. & Roth, B. J. Cisplatin in the Management of Breast Cancer. *Seminars in oncology* **16,** 110–115 (Aug. 1989).
17. Asleh, K., Riaz, N. & Nielsen, T. O. Heterogeneity of Triple Negative Breast Cancer: Current Advances in Subtyping and Treatment Implications. *Journal of Experimental & Clinical Cancer Research* **41,** 265. doi:10.1186/s13046-022-02476-1 (Sept. 2022).
18. Couch, F. J. *et al.* Inherited Mutations in 17 Breast Cancer Susceptibility Genes Among a Large Triple-Negative Breast Cancer Cohort Unselected for Family History of Breast Cancer. *Journal of Clinical Oncology* **33,** 304–311. doi:10.1200/JCO.2014.57.1414 (Feb. 2015).
19. Telli, M. L. *et al.* Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer. *Clinical Cancer Research* **22,** 3764–3773. doi:10.1158/1078-0432.CCR-15-2477 (Aug. 2016).
20. Cortazar, P. & Geyer, C. E. Pathological Complete Response in Neoadjuvant Treatment of Breast Cancer. *Annals of Surgical Oncology* **22,** 1441–1446. doi:10.1245/s10434-015-4404-8 (May 2015).

21. Lehmann, B. D. *et al.* Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies. *The Journal of Clinical Investigation* **121,** 2750–2767. doi:10.1172/JCI45014 (July 2011).

22. Lehmann, B. D. & Pietenpol, J. A. Identification and Use of Biomarkers in Treatment Strategies for Triple-Negative Breast Cancer Subtypes. *The Journal of Pathology* **232,** 142–150. doi:10.1002/path.4280 (2014).

23. Masuda, H. *et al.* Differential Response to Neoadjuvant Chemotherapy Among 7 Triple-Negative Breast Cancer Molecular Subtypes. *Clinical Cancer Research* **19,** 5533–5540. doi:10.1158/1078-0432.CCR-13-0799 (Oct. 2013).

24. Burstein, M. D. *et al.* Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer. *Clinical Cancer Research* **21,** 1688–1698. doi:10.1158/1078-0432.CCR-14-0432 (Apr. 2015).

25. Liu, Y.-R. *et al.* Comprehensive Transcriptome Analysis Identifies Novel Molecular Subtypes and Subtype-Specific RNAs of Triple-Negative Breast Cancer. *Breast Cancer Research* **18,** 33. doi:10.1186/s13058-016-0690-8 (Dec. 2016).

26. Lehmann, B. D. *et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLOS ONE* **11** (ed Sapino, A.) e0157368. doi:10.1371/journal.pone.0157368 (June 2016).

27. Costa, R. *et al.* Targeting Epidermal Growth Factor Receptor in Triple Negative Breast Cancer: New Discoveries and Practical Insights for Drug Development. *Cancer Treatment Reviews* **53,** 111–119. doi:10.1016/j.ctrv.2016.12.010 (Feb. 2017).

28. Kassam, F. *et al.* Survival Outcomes for Patients with Metastatic Triple-Negative Breast Cancer: Implications for Clinical Practice and Trial Design. *Clinical Breast Cancer* **9,** 29–33. doi:10.3816/CBC.2009.n.005 (Feb. 2009).

29. Staudacher, L. *et al.* Platinum-Based Chemotherapy in Metastatic Triple-Negative Breast Cancer: The Institut Curie Experience. *Annals of Oncology* **22,** 848–856. doi:10.1093/annonc/mdq461 (Apr. 2011).

30. Isakoff, S. J. *et al.* TBCRC009: A Multicenter Phase II Clinical Trial of Platinum Monotherapy With Biomarker Assessment in Metastatic Triple-Negative Breast Cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **33,** 1902–1909. doi:10.1200/JCO.2014.57.6660 (June 2015).

31. Loi, S. Tumor-Infiltrating Lymphocytes, Breast Cancer Subtypes and Therapeutic Efficacy. *Oncoimmunology* **2,** e24720. doi:10.4161/onci.24720 (July 2013).

32. Loi, S. *et al.* Tumor Infiltrating Lymphocytes Are Prognostic in Triple Negative Breast Cancer and Predictive for Trastuzumab Benefit in Early Breast Cancer: Results from the FinHER Trial. *Annals of Oncology: Official Journal of the European Society for Medical Oncology* **25,** 1544–1550. doi:10.1093/annonc/mdu112 (Aug. 2014).

33. Adams, S. *et al.* Prognostic Value of Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancers From Two Phase III Randomized Adjuvant Breast Cancer Trials: ECOG 2197 and ECOG 1199. *Journal of Clinical Oncology* **32,** 2959–2966. doi:10.1200/JCO.2013.55.0491 (Sept. 2014).

34. Ibrahim, E. M., Al-Foheidi, M. E., Al-Mansour, M. M. & Kazkaz, G. A. The Prognostic Value of Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: A Meta-Analysis. *Breast Cancer Research and Treatment* **148,** 467–476. doi:10.1007/s10549-014-3185-2 (Dec. 2014).

35. Loi, S. *et al.* Prognostic and Predictive Value of Tumor-Infiltrating Lymphocytes in a Phase III Randomized Adjuvant Breast Cancer Trial in Node-Positive Breast Cancer Comparing the Addition of Docetaxel to Doxorubicin with Doxorubicin-Based Chemotherapy: BIG 02-98. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **31,** 860–867. doi:10.1200/JCO.2011.41.0902 (Mar. 2013).

36. Ono, M. *et al.* Tumor-Infiltrating Lymphocytes Are Correlated with Response to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. *Breast Cancer Research and Treatment* **132,** 793–805. doi:10.1007/s10549-011-1554-7 (Apr. 2012).

37. Denkert, C. *et al.* Tumor-Infiltrating Lymphocytes and Response to Neoadjuvant Chemotherapy with or without Carboplatin in Human Epidermal Growth Factor Receptor 2-Positive and Triple-Negative Primary Breast Cancers. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **33,** 983–991. doi:10.1200/JCO.2014.58.1967 (Mar. 2015).

38. Alexandrov, L. B. *et al.* Signatures of Mutational Processes in Human Cancer. *Nature* **500,** 415–421. doi:10.1038/nature12477 (Aug. 2013).

39. Smid, M. *et al.* Patterns and Incidence of Chromosomal Instability and Their Prognostic Relevance in Breast Cancer Subtypes. *Breast Cancer Research and Treatment* **128,** 23–30. doi:10.1007/s10549-010-1026-5 (July 2011).

40. Castaneda, C. A. *et al.* Tumor Infiltrating Lymphocytes in Triple Negative Breast Cancer Receiving Neoadjuvant Chemotherapy. *World Journal of Clinical Oncology* **7,** 387–394. doi:10.5306/wjco.v7.i5.387 (Oct. 2016).

41. Beckers, R. K. *et al.* Programmed Death Ligand 1 Expression in Triple-Negative Breast Cancer Is Associated with Tumour-Infiltrating Lymphocytes and Improved Outcome. *Histopathology* **69,** 25–34. doi:10.1111/his.12904 (2016).

42. Schmid, P. *et al.* Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. *The New England Journal of Medicine* **379,** 2108–2121. doi:10.1056/NEJMoa1809615 (Nov. 2018).

43. Lev, S. Targeted Therapy and Drug Resistance in Triple-Negative Breast Cancer: The EGFR Axis. *Biochemical Society Transactions* **48,** 657–665. doi:10.1042/BST20191055 (Apr. 2020).

44. Pagliarini, R., Shao, W. & Sellers, W. R. Oncogene Addiction: Pathways of Therapeutic Response, Resistance, and Road Maps toward a Cure. *EMBO reports* **16,** 280–296. doi:10.15252/embr.201439949 (Mar. 2015).

45. Kazandjian, D. *et al.* FDA Approval of Gefitinib for the Treatment of Patients with Metastatic EGFR Mutation–Positive Non–Small Cell Lung Cancer. *Clinical Cancer Research* **22,** 1307–1312 (2016).

46. Fang, B. Development of Synthetic Lethality Anticancer Therapeutics. *Journal of Medicinal Chemistry* **57,** 7859–7873. doi:10.1021/jm500415t (Oct. 2014).

47. Brunen, D. & Bernards, R. Exploiting Synthetic Lethality to Improve Cancer Therapy. *Nature Reviews Clinical Oncology* **14,** 331–332. doi:10.1038/nrclinonc.2017.46 (June 2017).

48. Chatterjee, N. & Bivona, T. G. Polytherapy and Targeted Cancer Drug Resistance. *Trends in Cancer* **5,** 170–182. doi:10.1016/j.trecan.2019.02.003 (Mar. 2019).

49. Zong, Y. & Goldstein, A. S. Adaptation or Selection—Mechanisms of Castration-Resistant Prostate Cancer. *Nature Reviews Urology* **10,** 90–98. doi:10.1038/nrurol.2012.237 (Feb. 2013).

50. Chen, Y.-L. *et al.* Five Technologies for Detecting the EGFR T790M Mutation in the Circulating Cell-Free DNA of Patients with Non-Small Cell Lung Cancer: A Comparison. *Frontiers in Oncology* **9,** 631 (2019).

51. Byrne, K. M. *et al.* Bistability in the Rac1, PAK, and RhoA Signaling Network Drives Actin Cytoskeleton Dynamics and Cell Motility Switches. *Cell Systems* **2,** 38–48. doi:10.1016/j.cels.2016.01.003 (Jan. 2016).

52. Prahallad, A. & Bernards, R. Opportunities and Challenges Provided by Crosstalk between Signalling Pathways in Cancer. *Oncogene* **35,** 1073–1079. doi:10.1038/onc.2015.151 (Mar. 2016).

53. Shah, S. P. *et al.* The Clonal and Mutational Evolution Spectrum of Primary Triple-Negative Breast Cancers. *Nature* **486,** 395–399. doi:10.1038/nature10933 (June 2012).

54. Walerych, D., Napoli, M., Collavin, L. & Del Sal, G. The Rebel Angel: Mutant P53 as the Driving Oncogene in Breast Cancer. *Carcinogenesis* **33,** 2007–2017. doi:10.1093/carcin/bgs232 (Nov. 2012).

55. The Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* **490,** 61–70. doi:10.1038/nature11412 (Oct. 2012).

56. Turner, N. *et al.* Targeting Triple Negative Breast Cancer: Is P53 the Answer? *Cancer Treatment Reviews* **39,** 541–550. doi:10.1016/j.ctrv.2012.12.001 (Aug. 2013).

57. Synnott, N. *et al.* Mutant P53: A Novel Target for the Treatment of Patients with Triple-Negative Breast Cancer?: Mutant P53. *International Journal of Cancer* **140,** 234–246. doi:10.1002/ijc.30425 (Jan. 2017).

58. Venkitaraman, A. R. Cancer Suppression by the Chromosome Custodians, BRCA1 and BRCA2. *Science* **343,** 1470–1475. doi:10.1126/science.1252230 (Mar. 2014).

59. Couch, F. J., Nathanson, K. L. & Offit, K. Two Decades After BRCA: Setting Paradigms in Personalized Cancer Care and Prevention. *Science* **343,** 1466–1470. doi:10.1126/science.1251827 (Mar. 2014).

60. Lord, C. J. & Ashworth, A. BRCAness Revisited. *Nature Reviews Cancer* **16,** 110–120. doi:10.1038/nrc.2015.21 (Feb. 2016).

61. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in Sporadic Cancers. *Nature Reviews Cancer* **4,** 814–819. doi:10.1038/nrc1457 (Oct. 2004).

62. Prakash, R., Zhang, Y., Feng, W. & Jasin, M. Homologous Recombination and Human Health: The Roles of BRCA1, BRCA2, and Associated Proteins. *Cold Spring Harbor Perspectives in Biology* **7,** a016600. doi:10.1101/cshperspect.a016600 (Apr. 2015).

63. Byrum, A. K., Vindigni, A. & Mosammaparast, N. Defining and Modulating 'BRCAness'. *Trends in Cell Biology* **29,** 740–751. doi:10.1016/j.tcb.2019.06.005 (Sept. 2019).

64. Watkins, J. A., Irshad, S., Grigoriadis, A. & Tutt, A. N. J. Genomic Scars as Biomarkers of Homologous Recombination Deficiency and Drug Response in Breast and Ovarian Cancers. *Breast cancer research: BCR* **16,** 211. doi:10.1186/bcr3670 (June 2014).

65. Timms, K. M. *et al.* Association of BRCA1/2 Defects with Genomic Scores Predictive of DNA Damage Repair Deficiency among Breast Cancer Subtypes. *Breast cancer research: BCR* **16,** 475. doi:10.1186/s13058-014-0475-x (Dec. 2014).

66. Cruz, C. *et al.* RAD51 Foci as a Functional Biomarker of Homologous Recombination Repair and PARP Inhibitor Resistance in Germline BRCA-mutated Breast Cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology* **29,** 1203–1210. doi:10.1093/annonc/mdy099 (May 2018).

67. Fong, P. C. *et al.* Inhibition of Poly(ADP-ribose) Polymerase in Tumors from BRCA Mutation Carriers. *The New England Journal of Medicine* **361,** 123–134. doi:10.1056/NEJMoa0900212 (July 2009).

68. Tutt, A. *et al.* Oral Poly(ADP-ribose) Polymerase Inhibitor Olaparib in Patients with BRCA1 or BRCA2 Mutations and Advanced Breast Cancer: A Proof-of-Concept Trial. *Lancet (London, England)* **376,** 235–244. doi:10.1016/S0140-6736(10)60892-6 (July 2010).

69. Kaufman, B. *et al.* Olaparib Monotherapy in Patients with Advanced Cancer and a Germline BRCA1/2 Mutation. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **33,** 244–250. doi:10.1200/JCO.2014.56.2728 (Jan. 2015).

70. Robson, M. *et al.* Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *The New England Journal of Medicine* **377,** 523–533. doi:10.1056/NEJMoa1706450 (Aug. 2017).

71. Litton, J. K. *et al.* Talazoparib in Patients with Advanced Breast Cancer and a Germline BRCA Mutation. *The New England Journal of Medicine* **379,** 753–763. doi:10.1056/NEJMoa1802905 (Aug. 2018).

72. Exman, P., Barroso-Sousa, R. & Tolaney, S. M. Evidence to Date: Talazoparib in the Treatment of Breast Cancer. *OncoTargets and Therapy* **Volume 12,** 5177–5187. doi:10.2147/OTT.S184971 (July 2019).

73. Noordermeer, S. M. & van Attikum, H. PARP Inhibitor Resistance: A Tug-of-War in BRCA-Mutated Cells. *Trends in Cell Biology* **29,** 820–834. doi:10.1016/j.tcb.2019.07.008 (Oct. 2019).

74. Li, H. *et al.* PARP Inhibitor Resistance: The Underlying Mechanisms and Clinical Implications. *Molecular Cancer* **19,** 107. doi:10.1186/s12943-020-01227-0 (June 2020).

75. Jaspers, J. E. *et al.* Loss of 53BP1 Causes PARP Inhibitor Resistance in Brca1-Mutated Mouse Mammary Tumors. *Cancer Discovery* **3,** 68–81. doi:10.1158/2159-8290.CD-12-0049 (Jan. 2013).

76. Xu, G. *et al.* REV7 Counteracts DNA Double-Strand Break Resection and Affects PARP Inhibition. *Nature* **521,** 541–544. doi:10.1038/nature14328 (May 2015).

77. Patch, A.-M. *et al.* Whole–Genome Characterization of Chemoresistant Ovarian Cancer. *Nature* **521,** 489–494. doi:10.1038/nature14410 (May 2015).

78. Johnson, S. F. *et al.* CDK12 Inhibition Reverses De Novo and Acquired PARP Inhibitor Resistance in BRCA Wild-Type and Mutated Models of Triple-Negative Breast Cancer. *Cell Reports* **17,** 2367–2381. doi:10.1016/j.celrep.2016.10.077 (Nov. 2016).

79. Ning, J.-F. *et al.* Myc Targeted CDK18 Promotes ATR and Homologous Recombination to Mediate PARP Inhibitor Resistance in Glioblastoma. *Nature Communications* **10,** 2910. doi:10.1038/s41467-019-10993-5 (July 2019).

80. Liao, H., Ji, F., Helleday, T. & Ying, S. Mechanisms for Stalled Replication Fork Stabilization: New Targets for Synthetic Lethality Strategies in Cancer Treatments. *EMBO reports* **19,** e46263. doi:10.15252/embr.201846263 (Sept. 2018).

81. Rondinelli, B. *et al.* EZH2 Promotes Degradation of Stalled Replication Forks by Recruiting MUS81 through Histone H3 Trimethylation. *Nature Cell Biology* **19,** 1371–1378. doi:10.1038/ncb3626 (Nov. 2017).

82. Ray Chaudhuri, A. *et al.* Replication Fork Stability Confers Chemoresistance in BRCA-deficient Cells. *Nature* **535,** 382–387. doi:10.1038/nature18325 (July 2016).

83. Barber, L. J. *et al.* Secondary Mutations in BRCA2 Associated with Clinical Resistance to a PARP Inhibitor. *The Journal of Pathology* **229,** 422–429. doi:10.1002/path.4140 (2013).

84. Norquist, B. *et al.* Secondary Somatic Mutations Restoring BRCA1/2 Predict Chemotherapy Resistance in Hereditary Ovarian Carcinomas. *Journal of Clinical Oncology* **29,** 3008–3015. doi:10.1200/JCO.2010.34.2980 (Aug. 2011).

85. Domchek, . *et al.*. Abstract PD5-04: An Open-Label, Phase II Basket Study of Olaparib and Durvalumab (MEDIOLA): Updated Results in Patients with Germline BRCA-mutated (gBRCAm) Metastatic Breast Cancer (MBC). *Cancer Research* **79,** PD5–04. doi:10.1158/1538-7445.SABCS18-PD5-04 (Feb. 2019).

86. Nakagawa, M. *et al.* Expression of P53, Ki-67, E-cadherin, N-cadherin and TOP2A in Triple-Negative Breast Cancer. *Anticancer research* **31,** 2389–2393 (2011).

87. Gucalp, A. & Traina, T. A. Targeting the Androgen Receptor in Triple-Negative Breast Cancer. *Current Problems in Cancer* **40,** 141–150. doi:10.1016/j.currproblcancer.2016.09.004 (Mar. 2016).

88. Barton, V. N. *et al.* Androgen Receptor Biology in Triple Negative Breast Cancer: A Case for Classification as AR+ or Quadruple Negative Disease. *Hormones & cancer* **6,** 206–213. doi:10.1007/s12672-015-0232-3 (Dec. 2015).

89. Wang, C. *et al.* Prognostic Value of Androgen Receptor in Triple Negative Breast Cancer: A Meta-Analysis. *Oncotarget* **7,** 46482–46491. doi:10.18632/oncotarget.10208 (July 2016).

90. Maeda, T. *et al.* Immunohistochemical Co-Expression Status of Cytokeratin 5/6, Androgen Receptor, and P53 as Prognostic Factors of Adjuvant Chemotherapy for Triple Negative Breast Cancer. *Medical Molecular Morphology* **49,** 11–21. doi:10.1007/s00795-015-0109-0 (Mar. 2016).

91. Hilborn, E. *et al.* Androgen Receptor Expression Predicts Beneficial Tamoxifen Response in Oestrogen Receptor-α-Negative Breast Cancer. *British Journal of Cancer* **114,** 248–255. doi:10.1038/bjc.2015.464 (Feb. 2016).

92. Asano, Y. *et al.* Clinical Verification of Sensitivity to Preoperative Chemotherapy in Cases of Androgen Receptor-Expressing Positive Breast Cancer. *British Journal of Cancer* **114,** 14–20. doi:10.1038/bjc.2015.434 (Jan. 2016).

93. Luo, Y., Huang, W., Zhang, H. & Liu, G. Prognostic Significance of CD117 Expression and TP53 Missense Mutations in Triple-Negative Breast Cancer. *Oncology Letters.* doi:10.3892/ol.2018.8104 (Feb. 2018).

94. Lü, L. *et al.* MicroRNAs in the Prognosis of Triple-Negative Breast Cancer: A Systematic Review and Meta-Analysis. *Medicine* **96,** e7085. doi:10.1097/MD.0000000000007085 (June 2017).

95. Bao, C. *et al.* Exploring Specific Prognostic Biomarkers in Triple-Negative Breast Cancer. *Cell Death & Disease* **10,** 1–14. doi:10.1038/s41419-019-2043-x (Oct. 2019).

96. Malla, R. R. *et al.* A Perspective on the Diagnostics, Prognostics, and Therapeutics of microRNAs of Triple-Negative Breast Cancer. *Biophysical Reviews* **11,** 227–234. doi:10.1007/s12551-019-00503-8 (Apr. 2019).

97. Adams, B. D. *et al.* miR-34a Silences c-SRC to Attenuate Tumor Growth in Triple-Negative Breast Cancer. *Cancer Research* **76,** 927–939. doi:10.1158/0008-5472.CAN-15-2321 (Feb. 2016).

98. Zhang, K. *et al.* Long Non-Coding RNAs as Novel Biomarkers for Breast Cancer Invasion and Metastasis (Review). *Oncology Letters* **14,** 1895–1904. doi:10.3892/ol.2017.6462 (Aug. 2017).

99. Augoff, K., McCue, B., Plow, E. F. & Sossey-Alaoui, K. miR-31 and Its Host Gene lncRNA LOC554202 Are Regulated by Promoter Hypermethylation in Triple-Negative Breast Cancer. *Molecular Cancer* **11,** 5. doi:10.1186/1476-4598-11-5 (Jan. 2012).

100. Sakthianandeswaren, A., Liu, S. & Sieber, O. M. Long Noncoding RNA LINP1: Scaffolding Non-Homologous End Joining. *Cell Death Discovery* **2,** 1–2. doi:10.1038/cddiscovery.2016.59 (July 2016).

101. Shi, F., Xiao, F., Ding, P., Qin, H. & Huang, R. Long Noncoding RNA Highly Up-regulated in Liver Cancer Predicts Unfavorable Outcome and Regulates Metastasis by MMPs in Triple-negative Breast Cancer. *Archives of Medical Research* **47,** 446–453. doi:10.1016/j.arcmed.2016.11.001 (Aug. 2016).

102. Baselga, J. Targeting the Phosphoinositide-3 (PI3) Kinase Pathway in Breast Cancer. *The Oncologist* **16 Suppl 1,** 12–19. doi:10.1634/theoncologist.2011-S1-12 (2011).

103. Porta, C., Paglino, C. & Mosca, A. Targeting PI3K/Akt/mTOR Signaling in Cancer. *Frontiers in Oncology* **4** (2014).

104. Pascual, J. & Turner, N. C. Targeting the PI3-kinase Pathway in Triple-Negative Breast Cancer. *Annals of Oncology* **30,** 1051–1060. doi:10.1093/annonc/mdz133 (July 2019).

105. Kim, S.-B. *et al.* Ipatasertib plus Paclitaxel versus Placebo plus Paclitaxel as First-Line Therapy for Metastatic Triple-Negative Breast Cancer (LOTUS): A Multicentre, Randomised, Double-Blind, Placebo-Controlled, Phase 2 Trial. *The Lancet Oncology* **18,** 1360–1372. doi:10.1016/S1470-2045(17)30450-3 (Oct. 2017).

106. Dent, R. *et al.* Overall Survival (OS) Update of the Double-Blind Placebo (PBO)-Controlled Randomized Phase 2 LOTUS Trial of First-Line Ipatasertib (IPAT) + Paclitaxel (PAC) for Locally Advanced/Metastatic Triple-Negative Breast Cancer (mTNBC). *Journal of Clinical Oncology* **36,** 1008–1008. doi:10.1200/JCO.2018.36.15_suppl.1008 (May 2018).

107. Ribatti, D. *et al.* Angiogenesis and Antiangiogenesis in Triple-Negative Breast Cancer. *Translational Oncology* **9,** 453–457. doi:10.1016/j.tranon.2016.07.002 (Oct. 2016).

108. Linderholm, B. *et al.* Significantly Higher Levels of Vascular Endothelial Growth Factor (VEGF) and Shorter Survival Times for Patients with Primary Operable Triple-Negative Breast Cancer. *Annals of Oncology* **20,** 1639–1646. doi:10.1093/annonc/mdp062 (Oct. 2009).

109. Chanana, P. *et al.* Significance of Serum Vascular Endothelial Growth Factor and Cancer Antigen 15.3 in Patients with Triple Negative Breast Cancer. *Journal of Radiotherapy in Practice* **13,** 60–67. doi:10.1017/S146039691200057X (Mar. 2014).

110. Cameron, D. *et al.* Adjuvant Bevacizumab-Containing Therapy in Triple-Negative Breast Cancer (BEATRICE): Primary Results of a Randomised, Phase 3 Trial. *The Lancet Oncology* **14,** 933–942. doi:10.1016/S1470-2045(13)70335-8 (Sept. 2013).

111. Sigismund, S., Avanzato, D. & Lanzetti, L. Emerging Functions of the EGFR in Cancer. *Molecular Oncology* **12,** 3–20. doi:10.1002/1878-0261.12155 (2018).

112. Graus-Porta, D., Beerli, R. R., Daly, J. M. & Hynes, N. E. ErbB-2, the Preferred Heterodimerization Partner of All ErbB Receptors, Is a Mediator of Lateral Signaling. *The EMBO journal* **16,** 1647–1655. doi:10.1093/emboj/16.7.1647 (Apr. 1997).

113. Lemmon, M. A., Schlessinger, J. & Ferguson, K. M. The EGFR Family: Not So Prototypical Receptor Tyrosine Kinases. *Cold Spring Harbor Perspectives in Biology* **6,** a020768. doi:10.1101/cshperspect.a020768 (Apr. 2014).

114. Waterman, H. *et al.* A Mutant EGF-receptor Defective in Ubiquitylation and Endocytosis Unveils a Role for Grb2 in Negative Signaling. *The EMBO Journal* **21,** 303–313. doi:10.1093/emboj/21.3.303 (Feb. 2002).

115. Prenzel, N., Zwick, E., Leserer, M. & Ullrich, A. Tyrosine Kinase Signalling in Breast Cancer: Epidermal Growth Factor Receptor - Convergence Point for Signal Integration and Diversification. *Breast Cancer Research* **2,** 184–190. doi:10.1186/bcr52 (2000).

116. Wang, Z. Transactivation of Epidermal Growth Factor Receptor by G Protein-Coupled Receptors: Recent Progress, Challenges and Future Research. *International Journal of Molecular Sciences* **17,** 95. doi:10.3390/ijms17010095 (Jan. 2016).

117. Davis, N. M. *et al.* Deregulation of the EGFR/PI3K/PTEN/Akt/mTORC1 Pathway in Breast Cancer: Possibilities for Therapeutic Intervention. *Oncotarget* **5,** 4603–4650. doi:10.18632/oncotarget.2209 (July 2014).

118. Brand, T. M., Iida, M., Li, C. & Wheeler, D. L. The Nuclear Epidermal Growth Factor Receptor Signaling Network and Its Role in Cancer. *Discovery Medicine* **12,** 419–432 (Nov. 2011).

119. Huang, W.-C. *et al.* Nuclear Translocation of Epidermal Growth Factor Receptor by Akt-dependent Phosphorylation Enhances Breast Cancer-resistant Protein Expression in Gefitinib-resistant Cells*. *Journal of Biological Chemistry* **286,** 20558–20568. doi:10.1074/jbc.M111.240796 (June 2011).

120. Lo, H.-W. *et al.* Nuclear Interaction of EGFR and STAT3 in the Activation of the iNOS/NO Pathway. *Cancer Cell* **7,** 575–589. doi:10.1016/j.ccr.2005.05.007 (June 2005).

121. Park, H. S. *et al.* High EGFR Gene Copy Number Predicts Poor Outcome in Triple-Negative Breast Cancer. *Modern Pathology* **27,** 1212–1222. doi:10.1038/modpathol.2013.251 (Sept. 2014).

122. Bhargava, R. *et al.* EGFR Gene Amplification in Breast Cancer: Correlation with Epidermal Growth Factor Receptor mRNA and Protein Expression and HER-2 Status and Absence of EGFR-activating Mutations. *Modern Pathology* **18,** 1027–1033. doi:10.1038/modpathol.3800438 (Aug. 2005).

123. Teng, Y. H.-F. *et al.* Mutations in the Epidermal Growth Factor Receptor (EGFR) Gene in Triple Negative Breast Cancer: Possible Implications for Targeted Therapy. *Breast cancer research: BCR* **13,** R35. doi:10.1186/bcr2857 (Apr. 2011).

124. Ueno, N. T. & Zhang, D. Targeting EGFR in Triple Negative Breast Cancer. *Journal of Cancer* **2,** 324–328. doi:10.7150/jca.2.324 (2011).

125. Corkery, B., Crown, J., Clynes, M. & O'Donovan, N. Epidermal Growth Factor Receptor as a Potential Therapeutic Target in Triple-Negative Breast Cancer. *Annals of Oncology* **20,** 862–867. doi:10.1093/annonc/mdn710 (May 2009).

126. Nakai, K., Hung, M.-C. & Yamaguchi, H. A Perspective on Anti-EGFR Therapies Targeting Triple-Negative Breast Cancer. *American Journal of Cancer Research* **6,** 1609–1623 (Aug. 2016).

127. Wang, D.-Y., Jiang, Z., Ben-David, Y., Woodgett, J. R. & Zacksenhaus, E. Molecular Stratification within Triple-Negative Breast Cancer Subtypes. *Scientific Reports* **9,** 19107. doi:10.1038/s41598-019-55710-w (Dec. 2019).

128. Wander, S. A., Hennessy, B. T. & Slingerland, J. M. Next-Generation mTOR Inhibitors in Clinical Oncology: How Pathway Complexity Informs Therapeutic Strategy. *The Journal of Clinical Investigation* **121,** 1231–1241. doi:10.1172/JCI44145 (Apr. 2011).

129. Maira, S.-M. *et al.* Identification and Characterization of NVP-BEZ235, a New Orally Available Dual Phosphatidylinositol 3-Kinase/Mammalian Target of Rapamycin Inhibitor with Potent in Vivo Antitumor Activity. *Molecular Cancer Therapeutics* **7,** 1851–1863. doi:10.1158/1535-7163.MCT-08-0017 (July 2008).

130. Balak, M. N. *et al.* Novel D761Y and Common Secondary T790M Mutations in Epidermal Growth Factor Receptor–Mutant Lung Adenocarcinomas with Acquired Resistance to Kinase Inhibitors. *Clinical Cancer Research* **12,** 6494–6501. doi:10.1158/1078-0432.CCR-06-1570 (Nov. 2006).

131. Yoshida, T., Zhang, G. & Haura, E. B. Targeting Epidermal Growth Factor Receptor: Central Signaling Kinase in Lung Cancer. *Biochemical Pharmacology. Targeted Cancer Therapy* **80,** 613–623. doi:10.1016/j.bcp.2010.05.014 (Sept. 2010).

132. Bonner, J. A. *et al.* Radiotherapy plus Cetuximab for Locoregionally Advanced Head and Neck Cancer: 5-Year Survival Data from a Phase 3 Randomised Trial, and Relation between Cetuximab-Induced Rash and Survival. *The Lancet Oncology* **11,** 21–28. doi:10.1016/S1470-2045(09)70311-0 (Jan. 2010).

133. Heinemann, V. *et al.* FOLFIRI plus Cetuximab versus FOLFIRI plus Bevacizumab as First-Line Treatment for Patients with Metastatic Colorectal Cancer (FIRE-3): A Randomised, Open-Label, Phase 3 Trial. *The Lancet Oncology* **15,** 1065–1075. doi:10.1016/S1470-2045(14)70330-4 (Sept. 2014).

134. Pirker, R. *et al.* Cetuximab plus Chemotherapy in Patients with Advanced Non-Small-Cell Lung Cancer (FLEX): An Open-Label Randomised Phase III Trial. *The Lancet* **373,** 1525–1531. doi:10.1016/S0140-6736(09)60569-9 (May 2009).

135. Bean, J. *et al.* MET Amplification Occurs with or without T790M Mutations in EGFR Mutant Lung Tumors with Acquired Resistance to Gefitinib or Erlotinib. *Proceedings of the National Academy of Sciences* **104,** 20932–20937. doi:10.1073/pnas.0710370104 (Dec. 2007).

136. Engelman, J. A. *et al.* MET Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling. *Science* **316,** 1039–1043. doi:10.1126/science.1141478 (May 2007).

137. Yi, Y. W. *et al.* Dual Inhibition of EGFR and MET Induces Synthetic Lethality in Triple-Negative Breast Cancer Cells through Downregulation of Ribosomal Protein S6. *International Journal of Oncology* **47,** 122–132. doi:10.3892/ijo.2015.2982 (July 2015).

138. Linklater, E. S. *et al.* Targeting MET and EGFR Crosstalk Signaling in Triple-Negative Breast Cancers. *Oncotarget* **7,** 69903–69915. doi:10.18632/oncotarget.12065 (Oct. 2016).

139. Kim, Y. J. *et al.* MET Is a Potential Target for Use in Combination Therapy with EGFR Inhibition in Triple-Negative/Basal-like Breast Cancer: MET/EGFR as Potential Targets in TNBC/BLBC. *International Journal of Cancer* **134,** 2424–2436. doi:10.1002/ijc.28566 (May 2014).

140. Zhang, Z. *et al.* Activation of the AXL Kinase Causes Resistance to EGFR-targeted Therapy in Lung Cancer. *Nature Genetics* **44,** 852–860. doi:10.1038/ng.2330 (Aug. 2012).

141. Meyer, A. S., Miller, M. A., Gertler, F. B. & Lauffenburger, D. A. The Receptor AXL Diversifies EGFR Signaling and Limits the Response to EGFR-Targeted Inhibitors in Triple-Negative Breast Cancer Cells. *Science Signaling* **6,** ra66–ra66. doi:10.1126/scisignal.2004155 (Aug. 2013).

142. Craig, D. *et al.* Genome and Transcriptome Sequencing in Prospective Metastatic Triple-Negative Breast Cancer Uncovers Therapeutic Vulnerabilities. *Molecular Cancer Therapeutics* **12,** 104–116. doi:10.1158/1535-7163.MCT-12-0781 (2013).

143. Mirzoeva, O. *et al.* Basal Subtype and MAPK/ERK Kinase (MEK)-Phosphoinositide 3-Kinase Feedback Signaling Determine Susceptibility of Breast Cancer Cells to MEK Inhibition. *Cancer Research* **69,** 565–572. doi:10.1158/0008-5472.CAN-08-3389 (2009).

144. Maiello, M. R. *et al.* EGFR and MEK Blockade in Triple Negative Breast Cancer Cells. *Journal of Cellular Biochemistry* **116,** 2778–2785. doi:10.1002/jcb.25220 (2015).

145. Liu, T. *et al.* Combinatorial Effects of Lapatinib and Rapamycin in Triple-Negative Breast Cancer Cells. *Molecular Cancer Therapeutics* **10,** 1460–1469. doi:10.1158/1535-7163.MCT-10-0925 (2011).

146. Tao, J. J. *et al.* Antagonism of EGFR and HER3 Enhances the Response to Inhibitors of the PI3K-Akt Pathway in Triple-Negative Breast Cancer. *Science Signaling* **7,** ra29–ra29. doi:10.1126/scisignal.2005125 (Mar. 2014).

147. Verma, N. *et al.* Targeting of PYK2 Synergizes with EGFR Antagonists in Basal-like TNBC and Circumvents HER3-Associated Resistance via the NEDD4–NDRG1 Axis. *Cancer Research* **77,** 86–99. doi:10.1158/0008-5472.CAN-16-1797 (Jan. 2017).

148. Bae, D.-H. *et al.* The Role of NDRG1 in the Pathology and Potential Treatment of Human Cancers. *Journal of Clinical Pathology* **66,** 911–917. doi:10.1136/jclinpath-2013-201692 (Nov. 2013).

149. Shin, S.-Y., Müller, A.-K., Verma, N., Lev, S. & Nguyen, L. K. Systems Modelling of the EGFR-PYK2-c-Met Interaction Network Predicts and Prioritizes Synergistic Drug Combinations for Triple-Negative Breast Cancer. *PLOS Computational Biology* **14** (ed Ma'ayan, A.) e1006192. doi:10.1371/journal.pcbi.1006192 (June 2018).

150. Nguyen, L. K. Dynamics of Ubiquitin-Mediated Signalling: Insights from Mathematical Modelling and Experimental Studies. *Briefings in Bioinformatics* **17,** 479–493. doi:10.1093/bib/bbv052 (May 2016).

151. Nguyen, L. K. & Kholodenko, B. N. Feedback Regulation in Cell Signalling: Lessons for Cancer Therapeutics. *Seminars in Cell & Developmental Biology. Gap Junctions* **50,** 85–94. doi:10.1016/j.semcdb.2015.09.024 (Feb. 2016).

152. Finn, R. *et al.* Estrogen Receptor, Progesterone Receptor, Human Epidermal Growth Factor Receptor 2 (HER2), and Epidermal Growth Factor Receptor Expression and Benefit from Lapatinib in a Randomized Trial of Paclitaxel with Lapatinib or Placebo as First-Line Treatment in HER2-negative or Unknown Metastatic Breast Cancer. *Journal of Clinical Oncology* **27,** 3908–3915. doi:10.1200/JCO.2008.18.1925 (2009).

153. Schuler, M. *et al.* A Phase II Trial to Assess Efficacy and Safety of Afatinib in Extensively Pretreated Patients with HER2-negative Metastatic Breast Cancer. *Breast Cancer Research and Treatment* **134,** 1149–1159. doi:10.1007/s10549-012-2126-1 (Aug. 2012).

154. Mullooly, M. *et al.* Neratinib to Inhibit the Growth of Triple-Negative Breast Cancer Cells. *Journal of Clinical Oncology* **33,** 1099–1099. doi:10.1200/jco.2015.33.15_suppl.1099 (May 2015).

155. Maurer, M. *et al.* Abstract OT1-4-03: Phase I Study of Erlotinib and Metformin in Triple Negative Breast Cancer. *Cancer Research* **73,** OT1-4–03. doi:10.1158/0008-5472.SABCS13-OT1-4-03 (Dec. 2013).

156. Nowsheen, S., Cooper, T., Stanley, J. A. & Yang, E. S. Synthetic Lethal Interactions between EGFR and PARP Inhibition in Human Triple Negative Breast Cancer Cells. *PLOS ONE* **7,** e46614. doi:10.1371/journal.pone.0046614 (Oct. 2012).

157. Soliman, H., Khalil, F. & Antonia, S. PD-L1 Expression Is Increased in a Subset of Basal Type Breast Cancer Cells. *PLOS ONE* **9,** e88557. doi:10.1371/journal.pone.0088557 (Feb. 2014).

158. Millis, S. Z., Feldman, R. A. & O'Shaughnessy, J. Genomic and Protein Alterations in 126 Triple-Negative (TN) Metaplastic Breast Cancers. *Journal of Clinical Oncology* **32,** 1029–1029. doi:10.1200/jco.2014.32.15_suppl.1029 (May 2014).

159. Nabholtz, J. M. *et al.* Multicentric Neoadjuvant Phase II Study of Panitumumab Combined with an Anthracycline/Taxane-Based Chemotherapy in Operable Triple-Negative Breast Cancer: Identification of Biologically Defined Signatures Predicting Treatment Impact. *Annals of Oncology* **25,** 1570–1577. doi:10.1093/annonc/mdu183 (Aug. 2014).

160. Wang, W.-M. *et al.* Epidermal Growth Factor Receptor Inhibition Reduces Angiogenesis via Hypoxia-Inducible Factor-1α and Notch1 in Head Neck Squamous Cell Carcinoma. *PLOS ONE* **10,** e0119723. doi:10.1371/journal.pone.0119723 (Feb. 2015).

161. Patel, D. *et al.* Anti-Epidermal Growth Factor Receptor Monoclonal Antibody Cetuximab Inhibits EGFR/HER-2 Heterodimerization and Activation. *International Journal of Oncology* **34,** 25–32 (Jan. 2009).

162. Kimura, H. *et al.* Antibody-Dependent Cellular Cytotoxicity of Cetuximab against Tumor Cells with Wild-Type or Mutant Epidermal Growth Factor Receptor. *Cancer Science* **98,** 1275–1280. doi:10.1111/j.1349-7006.2007.00510.x (2007).

163. Srivastava, R. M. *et al.* Cetuximab-Activated Natural Killer (NK) and Dendritic Cells (DC) Collaborate to Trigger Tumor Antigen-Specific T Cell Immunity in Head and Neck Cancer Patients. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19,** 1858–1872. doi:10.1158/1078-0432.CCR-12-2426 (Apr. 2013).

164. Correale, P. *et al.* Cetuximab ± Chemotherapy Enhances Dendritic Cell-Mediated Phagocytosis of Colon Cancer Cells and Ignites a Highly Efficient Colon Cancer Antigen-Specific Cytotoxic T-cell Response in Vitro. *International Journal of Cancer* **130,** 1577–1589. doi:10.1002/ijc.26181 (2012).

165. Hsu, Y.-F. *et al.* Complement Activation Mediates Cetuximab Inhibition of Non-Small Cell Lung Cancer Tumor Growth in Vivo. *Molecular Cancer* **9,** 139. doi:10.1186/1476-4598-9-139 (June 2010).

166. Dechant, M. *et al.* Complement-Dependent Tumor Cell Lysis Triggered by Combinations of Epidermal Growth Factor Receptor Antibodies. *Cancer Research* **68,** 4998–5003. doi:10.1158/0008-5472.CAN-07-6226 (2008).

167. Prewett, M. *et al.* Enhanced Antitumor Activity of Anti-Epidermal Growth Factor Receptor Monoclonal Antibody IMC-C225 in Combination with Irinotecan (CPT-11) against Human Colorectal Tumor Xenografts. *Clinical Cancer Research* **8,** 994–1003 (2002).

168. Brand, T. M. *et al.* Nuclear Epidermal Growth Factor Receptor Is a Functional Molecular Target in Triple-Negative Breast Cancer. *Molecular cancer therapeutics* **13,** 1356–1368. doi:10.1158/1535-7163.MCT-13-1021 (May 2014).

169. Baselga, J. *et al.* Randomized Phase II Study of the Anti-Epidermal Growth Factor Receptor Monoclonal Antibody Cetuximab with Cisplatin versus Cisplatin Alone in Patients with Metastatic Triple-Negative Breast Cancer. *Journal of Clinical Oncology* **31,** 2586–2592. doi:10.1200/JCO.2012.46.2408 (2013).

170. Carey, L. A. *et al.* TBCRC 001: Randomized Phase II Study of Cetuximab in Combination With Carboplatin in Stage IV Triple-Negative Breast Cancer. *Journal of Clinical Oncology* **30,** 2615–2623. doi:10.1200/JCO.2010.34.5579 (July 2012).

171. Cowherd, S. *et al.* A Phase II Clinical Trial of Weekly Paclitaxel and Carboplatin in Combination with Panitumumab in Metastatic Triple Negative Breast Cancer. *Cancer Biology & Therapy* **16,** 678–683. doi:10.1080/15384047.2015.1026481 (May 2015).

172. Friedman, L. M. *et al.* Synergistic Down-Regulation of Receptor Tyrosine Kinases by Combinations of mAbs: Implications for Cancer Immunotherapy. *Proceedings of the National Academy of Sciences* **102,** 1915–1920. doi:10.1073/pnas.0409610102 (Feb. 2005).

173. Spangler, J. B. *et al.* Combination Antibody Treatment Down-Regulates Epidermal Growth Factor Receptor by Inhibiting Endosomal Recycling. *Proceedings of the National Academy of Sciences* **107,** 13252–13257. doi:10.1073/pnas.0913476107 (July 2010).

174. Ferraro, D. A. *et al.* Inhibition of Triple-Negative Breast Cancer Models by Combinations of Antibodies to EGFR. *Proceedings of the National Academy of Sciences* **110,** 1815–1820. doi:10.1073/pnas.1220763110 (Jan. 2013).

175. Leal, M. *et al.* Antibody–Drug Conjugates: An Emerging Modality for the Treatment of Cancer. *Annals of the New York Academy of Sciences* **1321,** 41–54. doi:10.1111/nyas.12499 (2014).

176. Reilly, E. B. *et al.* Characterization of ABT-806, a Humanized Tumor-Specific Anti-EGFR Monoclonal Antibody. *Molecular Cancer Therapeutics* **14,** 1141–1151. doi:10.1158/1535-7163.MCT-14-0820 (May 2015).

177. Phillips, A. C. *et al.* ABT-414, an Antibody–Drug Conjugate Targeting a Tumor-Selective EGFR Epitope. *Molecular Cancer Therapeutics* **15,** 661–669. doi:10.1158/1535-7163.MCT-15-0901 (Apr. 2016).

178. Bai, X., Ni, J., Beretov, J., Graham, P. & Li, Y. Triple-Negative Breast Cancer Therapeutic Resistance: Where Is the Achilles' Heel? *Cancer Letters* **497,** 100–111. doi:10.1016/j.canlet.2020.10.016 (Jan. 2021).

179. Brady, R. & Enderling, H. Mathematical Models of Cancer: When to Predict Novel Therapies, and When Not To. *Bulletin of Mathematical Biology* **81,** 3722–3731. doi:10.1007/s11538-019-00640-x (Oct. 2019).

180. Barbolosi, D., Ciccolini, J., Lacarelle, B., Barlési, F. & André, N. Computational Oncology — Mathematical Modelling of Drug Regimens for Precision Medicine. *Nature Reviews Clinical Oncology* **13,** 242–254. doi:10.1038/nrclinonc.2015.204 (Apr. 2016).

181. Transtrum, M. K. & Qiu, P. Bridging Mechanistic and Phenomenological Models of Complex Biological Systems. *PLoS Computational Biology* **12.** doi:10.1371/journal.pcbi.1004915. arXiv: 1509.06278 (2016).

182. Lightbody, G. *et al.* Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application. *Briefings in Bioinformatics* **20,** 1795–1811. doi:10.1093/bib/bby051 (Sept. 2019).

183. Bertucci, F. *et al.* Prospective High-Throughput Genome Profiling of Advanced Cancers: Results of the PERMED-01 Clinical Trial. *Genome Medicine* **13,** 87. doi:10.1186/s13073-021-00897-9 (May 2021).

184. Garralda, E. *et al.* New Clinical Trial Designs in the Era of Precision Medicine. *Molecular Oncology* **13,** 549–557. doi:10.1002/1878-0261.12465 (2019).

185. Colomer, R. *et al.* When Should We Order a next Generation Sequencing Test in a Patient with Cancer? *eClinicalMedicine* **25.** doi:10.1016/j.eclinm.2020.100487 (Aug. 2020).

186. Mosele, F. *et al.* Recommendations for the Use of Next-Generation Sequencing (NGS) for Patients with Metastatic Cancers: A Report from the ESMO Precision Medicine Working Group. *Annals of Oncology* **31,** 1491–1505. doi:10.1016/j.annonc.2020.07.014 (Nov. 2020).

187. Liu, A. *et al.* From Expression Footprints to Causal Pathways: Contextualizing Large Signaling Networks with CARNIVAL. *npj Systems Biology and Applications* **5,** 40. doi:10.1038/s41540-019-0118-z (Dec. 2019).

188. Desmond-Hellmann, S. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* doi:10.17226/13284 (2011).

189. Karolak, A., Markov, D. A., McCawley, L. J. & Rejniak, K. A. Towards Personalized Computational Oncology: From Spatial Models of Tumour Spheroids, to Organoids, to Tissues. *Journal of The Royal Society Interface* **15,** 20170703. doi:10.1098/rsif.2017.0703 (Jan. 2018).

190. Arnedos, M. *et al.* Precision Medicine for Metastatic Breast Cancer—Limitations and Solutions. *Nature Reviews Clinical Oncology* **12,** 693–704. doi:10.1038/nrclinonc.2015.123 (Dec. 2015).

191. Vargas, A. J. & Harris, C. C. Biomarker Development in the Precision Medicine Era: Lung Cancer as a Case Study. *Nature Reviews Cancer* **16,** 525–537. doi:10.1038/nrc.2016.56 (Aug. 2016).

192. Harbeck, N., Salem, M., Nitz, U., Gluz, O. & Liedtke, C. Personalized Treatment of Early-Stage Breast Cancer: Present Concepts and Future Directions. *Cancer Treatment Reviews* **36,** 584–594. doi:10.1016/j.ctrv.2010.04.007 (Dec. 2010).

193. Ozik, J. *et al.* High-Throughput Cancer Hypothesis Testing with an Integrated PhysiCell-EMEWS Workflow. *BMC Bioinformatics* **19,** 483. doi:10.1186/s12859-018-2510-x (Dec. 2018).

194. Ozik, J., Collier, N., Heiland, R., An, G. & Macklin, P. Learning-Accelerated Discovery of Immune-Tumour Interactions. *Molecular Systems Design & Engineering* **4,** 747–760. doi:10.1039/C9ME00036D (2019).

195. Hernandez-Boussard, T. *et al.* Digital Twins for Predictive Oncology Will Be a Paradigm Shift for Precision Cancer Care. *Nature Medicine* **27,** 2065–2066. doi:10.1038/s41591-021-01558-5 (Dec. 2021).

196. Fertig, E. J., Jaffee, E. M., Macklin, P., Stearns, V. & Wang, C. Forecasting Cancer: From Precision to Predictive Medicine. *Med* **2,** 1004–1010. doi:10.1016/j.medj.2021.08.007 (Sept. 2021).

197. Stahlberg, E. A. *et al.* Exploring Approaches for Predictive Cancer Patient Digital Twins: Opportunities for Collaboration and Innovation. *Frontiers in Digital Health* **4,** 1007784. doi:10.3389/fdgth.2022.1007784 (Oct. 2022).

198. Turner, N. C. & Reis-Filho, J. S. Genetic Heterogeneity and Cancer Drug Resistance. *The Lancet Oncology* **13,** e178–e185. doi:10.1016/S1470-2045(11)70335-7. PMID: 22469128 (Apr. 2012).

199. Shackleton, M., Quintana, E., Fearon, E. R. & Morrison, S. J. Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution. *Cell* **138,** 822–829. doi:10.1016/j.cell.2009.08.017 (Sept. 2009).

200. Barker, H. E., Paget, J. T., Khan, A. A. & Harrington, K. J. The Tumour Microenvironment after Radiotherapy: Mechanisms of Resistance and Recurrence. *Nature Reviews Cancer* **15,** 409–425. doi:10.1038/nrc3958. arXiv: 1507.02142v2 (2015).

201. Eyler, C. E. & Rich, J. N. Survival of the Fittest: Cancer Stem Cells in Therapeutic Resistance and Angiogenesis. *Journal of Clinical Oncology* **26,** 2839–2845. doi:10.1200/JCO.2007.15.1829 (June 2008).

202. Gatenby, R. A., Silva, A. S., Gillies, R. J. & Frieden, B. R. Adaptive Therapy. *Cancer Research* **69,** 4894–4903. doi:10.1158/0008-5472.CAN-08-3658 (June 2009).

203. Willyard, C. Cancer Therapy: An Evolved Approach. *Nature* **532,** 166–168. doi:10.1038/532166a (Apr. 2016).

204. Sottoriva, A. *et al.* A Big Bang Model of Human Colorectal Tumor Growth. *Nature Genetics* **47,** 209–216. doi:10.1038/ng.3214 (Mar. 2015).

205. Agur, Z., Elishmereni, M. & Kheifetz, Y. Personalizing Oncology Treatments by Predicting Drug Efficacy, Side-Effects, and Improved Therapy: Mathematics, Statistics, and Their Integration. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **6,** 239–253. doi:10.1002/wsbm.1263 (May 2014).

206. Wang, Z., Butner, J. D., Kerketta, R., Cristini, V. & Deisboeck, T. S. Simulating Cancer Growth with Multiscale Agent-Based Modeling. *Seminars in Cancer Biology* **30,** 70–78. doi:10.1016/j.semcancer.2014.04.001 (2015).

207. Schwab, J. D., Kühlwein, S. D., Ikonomi, N., Kühl, M. & Kestler, H. A. Concepts in Boolean Network Modeling: What Do They All Mean? *Computational and Structural Biotechnology Journal* **18,** 571–582. doi:10.1016/j.csbj.2020.03.001 (Jan. 2020).

208. Glass, L. & Kauffman, S. A. The Logical Analysis of Continuous, Non-Linear Biochemical Control Networks. *Journal of Theoretical Biology* **39,** 103–129. doi:10.1016/0022-5193(73)90208-7 (Apr. 1973).

209. Thomas, R. Regulatory Networks Seen as Asynchronous Automata: A Logical Description. *Journal of Theoretical Biology* **153,** 1–23. doi:10.1016/S0022-5193(05)80350-9 (Nov. 1991).

210. Harvey, I. & Bossomaier, T. *Time out of Joint: Attractors in Asynchronous* in *Fourth European Conf. on Artificial Life* **4** (1997), 67.

211. Saadatpour, A., Albert, I. & Albert, R. Attractor Analysis of Asynchronous Boolean Models of Signal Transduction Networks. *Journal of Theoretical Biology* **266,** 641–656. doi:10.1016/j.jtbi.2010.07.022 (Oct. 2010).

212. Bornholdt, S. Boolean Network Models of Cellular Regulation: Prospects and Limitations. *Journal of the Royal Society Interface* **5,** S85–S94. doi:10.1098/rsif.2008.0132.focus (2008).

213. Brandman, O., Ferrell, J. E., Li, R. & Meyer, T. Interlinked Fast and Slow Positive Feedback Loops Drive Reliable Cell Decisions. *Science* **310,** 496–498. doi:10.1126/science.1113834 (Oct. 2005).

214. Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus Asynchronous Modeling of Gene Regulatory Networks. *Bioinformatics (Oxford, England)* **24,** 1917–1925. doi:10.1093/bioinformatics/btn336 (Sept. 2008).

215. Trinh, V.-G., Hiraishi, K. & Benhamou, B. *Computing Attractors of Large-Scale Asynchronous Boolean Networks Using Minimal Trap Spaces* in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Association for Computing Machinery, New York, NY, USA, Aug. 2022), 1–10. doi:10.1145/3535508.3545520.

216. Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics (Oxford, England)* **18,** 261–274. doi:10.1093/bioinformatics/18.2.261 (Feb. 2002).

217. Monro, H. C. & Gaffney, E. A. Modelling Chemotherapy Resistance in Palliation and Failed Cure. *Journal of Theoretical Biology* **257,** 292–302. doi:10.1016/j.jtbi.2008.12.006 (Mar. 2009).

218. Powathil, G. G., Gordon, K. E., Hill, L. A. & Chaplain, M. A. J. Modelling the Effects of Cell-Cycle Heterogeneity on the Response of a Solid Tumour to Chemotherapy: Biological Insights from a Hybrid Multiscale Cellular Automaton Model. *Journal of Theoretical Biology* **308,** 1–19. doi:10.1016/j.jtbi.2012.05.015 (Sept. 2012).

219. Rejniak, K. A. *et al.* The Role of Tumor Tissue Architecture in Treatment Penetration and Efficacy: An Integrative Study. *Frontiers in Oncology* **3,** 1–13. doi:10.3389/fonc.2013.00111 (2013).

220. Lea, D. E. & Catcheside, D. G. The Mechanism of the Induction by Radiation of Chromosome Aberrations in Tradescantia. *Journal of Genetics* **44,** 216–245. doi:10.1007/BF02982830 (1942).

221. Jones, L., Hoban, P. & Metcalfe, P. The Use of the Linear Quadratic Model in Radiotherapy : *Australasian Physical & Engineering Sciences in Medicine* **24** (2001).

222. Jones, B. & Dale, R. G. in *Radiobiological Modelling in Radiation Oncology* (eds Dale, R. G. & Jones, B.) (The British Institute of Radiology, 2007).

223. Jones, B. & Dale, .. The Evolution of Practical Radiobiological Modelling. *The British Journal of Radiology,* 20180097. doi:10.1259/bjr.20180097 (Mar. 2018).

224. Jones, B. & Dale, R. G. Radiobiological Modeling and Clinical Trials. *International Journal of Radiation Oncology Biology Physics* **48,** 259–265. doi:10.1016/S0360-3016(00)00542-3 (2000).

225. Bull, J. A. & Byrne, H. M. The Hallmarks of Mathematical Oncology. *Proceedings of the IEEE* **110,** 523–540. doi:10.1109/JPROC.2021.3136715 (May 2022).

226. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70. doi:10.1016/s0092-8674(00)81683-9 (Jan. 2000).

227. Enderling, H. & A.J. Chaplain, M. Mathematical Modeling of Tumor Growth and Treatment. *Current Pharmaceutical Design* **20,** 4934–4940 (Sept. 2014).

228. Liu, J. *et al.* A Time-Resolved Experimental–Mathematical Model for Predicting the Response of Glioma Cells to Single-Dose Radiation Therapy. *Integrative Biology* **13,** 167–183. doi:10.1093/intbio/zyab010 (July 2021).

229. Paczkowski, M. *et al.* Reciprocal Interactions between Tumour Cell Populations Enhance Growth and Reduce Radiation Sensitivity in Prostate Cancer. *Communications Biology* **4,** 1–13. doi:10.1038/s42003-020-01529-5 (Jan. 2021).

230. Hahnfeldt, P., Panigrahy, D., Folkman, J. & Hlatky, L. Tumor Development under Angiogenic Signaling: A Dynamical Theory of Tumor Growth, Treatment Response, and Postvascular Dormancy1. *Cancer Research* **59,** 4770–4775 (Oct. 1999).

231. Sachs, R. K., Hlatky, L. R. & Hahnfeldt, P. Simple ODE Models of Tumor Growth and Anti-Angiogenic or Radiation Treatment. *Mathematical and Computer Modelling* **33,** 1297–1305. doi:10.1016/S0895-7177(00)00316-2 (June 2001).

232. Kuznetsov, V. A., Makalkin, I. A., Taylor, M. A. & Perelson, A. S. Nonlinear Dynamics of Immunogenic Tumors: Parameter Estimation and Global Bifurcation Analysis. *Bulletin of Mathematical Biology* **56,** 295–321. doi:10.1016/S0092-8240(05)80260-5 (Mar. 1994).

233. Butuc, I., Mirestean, C. & Iancu, D. A Nonlinear Model in the Dynamics of Tumour-Immune System Combined with Radiotherapy. *U.P.B. Sci. Bull.* (2019).

234. Ghaffari, A., Bahmaie, B. & Nazari, M. A Mixed Radiotherapy and Chemotherapy Model for Treatment of Cancer with Metastasis. *Mathematical Methods in the Applied Sciences* **39,** 4603–4617. doi:10.1002/mma.3887 (2016).

235. Al-Tuwairqi, S. M., Al-Johani, N. O. & Simbawa, E. A. Modeling Dynamics of Cancer Radiovirotherapy. *Journal of Theoretical Biology* **506,** 110405. doi:10.1016/j.jtbi.2020.110405 (Dec. 2020).

236. Greenspan, H. P. Models for the Growth of a Solid Tumor by Diffusion. *Studies in Applied Mathematics* **51,** 317–340. doi:10.1002/sapm1972514317 (1972).

237. Lewin, T. D., Maini, P. K., Moros, E. G., Enderling, H. & Byrne, H. M. The Evolution of Tumour Composition During Fractionated Radiotherapy: Implications for Outcome. *Bulletin of Mathematical Biology* **80,** 1207–1235. doi:10.1007/s11538-018-0391-9 (May 2018).

238. Swanson, K. R., Alvord Jr, E. C. & Murray, J. D. A Quantitative Model for Differential Motility of Gliomas in Grey and White Matter. *Cell Proliferation* **33,** 317–329. doi:10.1046/j.1365-2184.2000.00177.x (2000).

239. Swanson, K. R., Alvord, E. C. & Murray, J. D. Virtual Brain Tumours (Gliomas) Enhance the Reality of Medical Imaging and Highlight Inadequacies of Current Therapy. *British Journal of Cancer* **86,** 14–18. doi:10.1038/sj.bjc.6600021 (Jan. 2002).

240. Swanson, K. R. *et al.* Quantifying the Role of Angiogenesis in Malignant Progression of Gliomas: In Silico Modeling Integrates Imaging and Histology. *Cancer Research* **71,** 7366–7375. doi:10.1158/0008-5472.CAN-11-1399 (Dec. 2011).

241. Rockne, R. *et al.* Predicting the Efficacy of Radiotherapy in Individual Glioblastoma Patients in Vivo: A Mathematical Modeling Approach. *Physics in Medicine & Biology* **55,** 3271. doi:10.1088/0031-9155/55/12/001 (May 2010).

242. Hawkins-Daarud, A. *et al.* In Silico Analysis Suggests Differential Response to Bevacizumab and Radiation Combination Therapy in Newly Diagnosed Glioblastoma. *Journal of The Royal Society Interface* **12,** 20150388. doi:10.1098/rsif.2015.0388 (Aug. 2015).

243. Hubbard, M. E. & Byrne, H. M. Multiphase Modelling of Vascular Tumour Growth in Two Spatial Dimensions. *Journal of Theoretical Biology* **316,** 70–89. doi:10.1016/j.jtbi.2012.09.031 (Jan. 2013).

244. Lewin, T. D. *et al.* The Importance of Dead Material within a Tumour on the Dynamics in Response to Radiotherapy. *Physics in Medicine & Biology* **65,** 015007. doi:10.1088/1361-6560/ab4c27 (Jan. 2020).

245. Boemo, M. A. & Byrne, H. M. Mathematical Modelling of a Hypoxia-Regulated Oncolytic Virus Delivered by Tumour-Associated Macrophages. *Journal of Theoretical Biology* **461,** 102–116. doi:10.1016/j.jtbi.2018.10.044 (Jan. 2019).

246. Chen, C. Y., Byrne, H. M. & King, J. R. The Influence of Growth-Induced Stress from the Surrounding Medium on the Development of Multicell Spheroids. *Journal of Mathematical Biology* **43,** 191–220. doi:10.1007/s002850100091 (Sept. 2001).

247. Fritz, M., Lima, E. A. B. F., Nikolić, V., Oden, J. T. & Wohlmuth, B. Local and Nonlocal Phase-Field Models of Tumor Growth and Invasion Due to ECM Degradation. *Mathematical Models and Methods in Applied Sciences* **29,** 2433–2468. doi:10.1142/S0218202519500519 (Dec. 2019).

248. Vilanova, G., Colominas, I. & Gomez, H. A Mathematical Model of Tumour Angiogenesis: Growth, Regression and Regrowth. *Journal of The Royal Society Interface* **14,** 20160918. doi:10.1098/rsif.2016.0918 (Jan. 2017).

249. Graner, F. & Glazier, J. A. Simulation of Biological Cell Sorting Using a Two-Dimensional Extended Potts Model. *Physical Review Letters* **69,** 2013–2016. doi:10.1103/PhysRevLett.69.2013 (Sept. 1992).

250. Osborne, J. M., Fletcher, A. G., Pitt-Francis, J. M., Maini, P. K. & Gavaghan, D. J. Comparing Individual-Based Approaches to Modelling the Self-Organization of Multicellular Tissues. *PLOS Computational Biology* **13,** e1005387. doi:10.1371/journal.pcbi.1005387 (Feb. 2017).

251. Brüningk, S. C., Ziegenhein, P., Rivens, I., Oelfke, U. & ter Haar, G. A Cellular Automaton Model for Spheroid Response to Radiation and Hyperthermia Treatments. *Scientific Reports* **9,** 17674. doi:10.1038/s41598-019-54117-x (Nov. 2019).

252. Gao, X., McDonald, J. T., Hlatky, L. & Enderling, H. Acute and Fractionated Irradiation Differentially Modulate Glioma Stem Cell Division Kinetics. *Cancer Research* **73,** 1481–1490. doi:10.1158/0008-5472.CAN-12-3429 (Dec. 2012).

253. Kempf, H., Bleicher, M. & Meyer-Hermann, M. Spatio-Temporal Cell Dynamics in Tumour Spheroid Irradiation. *The European Physical Journal D* **60,** 177–193. doi:10.1140/epjd/e2010-00178-4 (Oct. 2010).

254. Wang, H., Ma, H., Sové, R. J., Emens, L. A. & Popel, A. S. Quantitative Systems Pharmacology Model Predictions for Efficacy of Atezolizumab and Nab-Paclitaxel in Triple-Negative Breast Cancer. *Journal for ImmunoTherapy of Cancer* **9,** e002100. doi:10.1136/jitc-2020-002100 (Feb. 2021).

255. Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. PhysiCell: An Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems. *PLOS Computational Biology* **14,** e1005991. doi:10.1371/journal.pcbi.1005991 (Feb. 2018).

256. Ghaffarizadeh, A., Friedman, S. H. & Macklin, P. BioFVM: An Efficient, Parallelized Diffusive Transport Solver for 3-D Biological Simulations. *Bioinformatics* **32,** 1256–1258. doi:10.1093/bioinformatics/btv730 (Apr. 2016).

257. Voukantsis, D., Kahn, K., Hadley, M., Wilson, R. & Buffa, F. M. Modeling Genotypes in Their Microenvironment to Predict Single- and Multi-Cellular Behavior. *GigaScience* **8.** doi:10.1093/gigascience/giz010 (Mar. 2019).

258. Cooper, F. R. *et al.* Chaste: Cancer, Heart and Soft Tissue Environment. *Journal of open source software* **5,** 1848. doi:10.21105/joss.01848 (Mar. 2020).

259. Bravo, R. R. *et al.* Hybrid Automata Library: A Flexible Platform for Hybrid Modeling with Real-Time Visualization. *PLOS Computational Biology* **16,** e1007635. doi:10.1371/journal.pcbi.1007635 (Mar. 2020).

260. Starruß, J., de Back, W., Brusch, L. & Deutsch, A. Morpheus: A User-Friendly Modeling Environment for Multiscale and Multicellular Systems Biology. *Bioinformatics* **30,** 1331–1332. doi:10.1093/bioinformatics/btt772 (May 2014).

261. Kang, S., Kahan, S., McDermott, J., Flann, N. & Shmulevich, I. Biocellion : Accelerating Computer Simulation of Multicellular Biological System Models. *Bioinformatics* **30,** 3101–3108. doi:10.1093/bioinformatics/btu498 (Nov. 2014).

262. Boujelben, A. *et al.* Multimodality Imaging and Mathematical Modelling of Drug Delivery to Glioblastomas. *Interface Focus* **6.** doi:10.1098/rsfs.2016.0039 (2016).

263. Karolak, A. *et al.* Targeting Ligand Specificity Linked to Tumor Tissue Topological Heterogeneity via Single-Cell Micro-Pharmacological Modeling. *Scientific Reports* **8,** 1–14. doi:10.1038/s41598-018-21883-z (2018).

264. Curtis, L. T., Wu, M., Lowengrub, J., Decuzzi, P. & Frieboes, H. B. Computational Modeling of Tumor Response to Drug Release from Vasculature-Bound Nanoparticles. *PLoS ONE* **10,** 1–17. doi:10.1371/journal.pone.0144888 (2015).

265. Owen, M. R. *et al.* Mathematical Modeling Predicts Synergistic Antitumor Effects of Combining a Macrophage-Based, Hypoxia-Targeted Gene Therapy with Chemotherapy. *Cancer Research* **71,** 2826–2837. doi:10.1158/0008-5472.CAN-10-2834 (Apr. 2011).

266. Leonard, F. *et al.* Enhanced Performance of Macrophage-Encapsulated Nanoparticle Albumin-Bound-Paclitaxel in Hypo-Perfused Cancer Lesions. *Nanoscale* **8,** 12544–12552. doi:10.1039/c5nr07796f (2016).

267. Fu, K. K. Biological Basis for the Interaction of Chemotherapeutic Agents and Radiation Therapy. *Cancer* **55,** 2123–30. doi:10.1002/1097-0142(19850501)55:9+{$<$}2123::AID-CNCR2820551415{$>$}3.0.CO;2-0. PMID: 3884135 (May 1985).

268. Hennequin, C. & Favaudon, V. Biological Basis for Chemo-Radiotherapy Interactions. *European Journal of Cancer* **38,** 223–230. doi:10.1016/S0959-8049(01)00360-4 (2002).

269. Pawlik, T. M. & Keyomarsi, K. Role of Cell Cycle in Mediating Sensitivity to Radiotherapy. *International Journal of Radiation Oncology Biology Physics* **59,** 928–942. doi:10.1016/j.ijrobp.2004.03.005 (2004).

270. Powathil, G., Kohandel, M., Sivaloganathan, S., Oza, A. & Milosevic, M. Mathematical Modeling of Brain Tumors: Effects of Radiotherapy and Chemotherapy. *Physics in Medicine and Biology* **52,** 3291–3306. doi:10.1088/0031-9155/52/11/023 (June 2007).

271. Powathil, G. G., Adamson, D. J. & Chaplain, M. A. Towards Predicting the Response of a Solid Tumour to Chemotherapy and Radiotherapy Treatments: Clinical Insights from a Computational Model. *PLoS Computational Biology* **9.** doi:10.1371/journal.pcbi.1003120 (2013).

272. Kohandel, M., Sivaloganathan, S. & Oza, A. Mathematical Modeling of Ovarian Cancer Treatments: Sequencing of Surgery and Chemotherapy. *Journal of Theoretical Biology* **242,** 62–68. doi:10.1016/j.jtbi.2006.02.001 (Sept. 2006).

273. MacKay, T. F., Stone, E. A. & Ayroles, J. F. The Genetics of Quantitative Traits: Challenges and Prospects. *Nature Reviews Genetics* **10,** 565–577. doi:10.1038/nrg2612. arXiv: 1011.1669v3 (2009).

274. Jiménez, J. I. Shedding Light on the Black Box Models of the Cell. *Microbial Biotechnology* **10,** 43–45. doi:10.1111/1751-7915.12489 (2017).

275. Yi, S. *et al.* Functional Variomics and Network Perturbation: Connecting Genotype to Phenotype in Cancer. *Nature Reviews Genetics* **18,** 395–410. doi:10.1038/nrg.2017.8 (2017).

276. Hanahan, D. & Coussens, L. M. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* **21,** 309–322. doi:10.1016/j.ccr.2012.02.022. arXiv: gr-qc/0208024 (2012).

277. Orgogozo, V., Morizot, B. & Martin, A. The Differential View of Genotype-Phenotype Relationships. *Frontiers in Genetics* **6,** 1–14. doi:10.3389/fgene.2015.00179 (May 2015).

278. Takiar, V., Ip, C. K., Gao, M., Mills, G. B. & Cheung, L. W. Neomorphic Mutations Create Therapeutic Challenges in Cancer. *Oncogene* **36,** 1607–1618. doi:10.1038/onc.2016.312 (2017).

279. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biology* **13.** doi:10.1371/journal.pbio.1002195 (2015).

280. National Cancer Institute. *The Cancer Genome Atlas* http://cancergenome.nih.gov/.

281. *International Cancer Genome Consortium* icgc.org.

282. *Library of Integrated Network-Based Cellular Signatures (LINCS)* http://www.lincsproject.org/LINCS/.

283. *Integrated Cancer Biology Program Search Form Search (ICBP)* http://icbp.mit.edu/integrated-cancer-biology-program.

284. *DepMap: The Cancer Dependency Map Project at Broad Institute* https://depmap.org/portal/.

285. *KEGG: Kyoto Encyclopedia of Genes and Genomes* https://www.genome.jp/kegg/.

286. Slenter, D. N. *et al.* WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Research* **46,** D661–D667. doi:10.1093/nar/gkx1064 (2018).

287. *Pathway Commons* https://www.pathwaycommons.org/.

288. Mills, G. B. An Emerging Toolkit for Targeted Cancer Therapies. *Genome Research* **22,** 177–182. doi:10.1101/gr.136044.111 (2012).

289. Werner, H. M. J., Mills, G. B. & Ram, P. T. Cancer Systems Biology: A Peek into the Future of Patient Care? *Nature Reviews Clinical Oncology* **11,** 167–176. doi:10.1038/nrclinonc.2014.6 (2014).

290. Chin, L., Hahn, W. C., Getz, G. & Meyerson, M. Making Sense of Cancer Genomic Data. *Genes and Development* **25,** 534–555. doi:10.1101/gad.2017311 (2011).

291. Derry, J. M. *et al.* Developing Predictive Molecular Maps of Human Disease through Community-Based Modeling. *Nature Genetics* **44,** 127–130. doi:10.1038/ng.1089. arXiv: 1306.3636v1 (2012).

292. Norman, T. C., Bountra, C., Edwards, A. M., Yamamoto, K. R. & Friend, S. H. Leveraging Crowdsourcing to Facilitate the Discovery of New Medicines. *Science Translational Medicine* **3,** 1–4. doi:10.1126/scitranslmed.3002678 (2011).

293. Visvader, J. E. & Stingl, J. Mammary Stem Cells and the Differentiation Hierarchy: Current Status and Perspectives. *Genes & Development* **28,** 1143–1158. doi:10.1101/gad.242511.114 (June 2014).

294. Schubert, M. *et al.* Perturbation-Response Genes Reveal Signaling Footprints in Cancer Gene Expression. *Nature Communications* **9,** 20. doi:10.1038/s41467-017-02391-6 (Jan. 2018).

295. Karaayvaz, M. *et al.* Unravelling Subclonal Heterogeneity and Aggressive Disease States in TNBC through Single-Cell RNA-seq. *Nature Communications* **9,** 3588. doi:10.1038/s41467-018-06052-0 (Sept. 2018).

296. *BRECO* https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/relationship-between-breast-cancer-cells-and-the-surrounding-tissues/.

297. Gambardella, G. *et al.* A Single-Cell Analysis of Breast Cancer Cell Lines to Study Tumour Heterogeneity and Drug Response. *Nature Communications* **13,** 1714. doi:10.1038/s41467-022-29358-6 (Mar. 2022).

298. Hyatt, D. C. & Ceresa, B. P. Cellular Localization of the Activated EGFR Determines Its Effect on Cell Growth in MDA-MB-468 Cells. *Experimental Cell Research* **314,** 3415–3425. doi:10.1016/j.yexcr.2008.08.020 (Nov. 2008).

299. Hao, Y. *et al.* Integrated Analysis of Multimodal Single-Cell Data. *Cell* **184,** 3573–3587.e29. doi:10.1016/j.cell.2021.04.048 (June 2021).

300. Tirosh, I. *et al.* Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-seq. *Science* **352,** 189–196. doi:10.1126/science.aad0501 (Apr. 2016).

301. Zappia, L. & Oshlack, A. Clustering Trees: A Visualization for Evaluating Clusterings at Multiple Resolutions. *GigaScience* **7.** doi:10.1093/gigascience/giy083 (July 2018).

302. Kuleshov, M. V. *et al.* Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Research* **44,** W90–W97. doi:10.1093/nar/gkw377 (July 2016).

303. Badia-i-Mompel, P. *et al.* decoupleR: Ensemble of Computational Methods to Infer Biological Activities from Omics Data. *Bioinformatics Advances* **2** (ed Kuijjer, M. L.) vbac016. doi:10.1093/bioadv/vbac016 (Jan. 2022).

304. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).

305. Blanco-Carmona, E. *Generating Publication Ready Visualizations for Single Cell Transcriptomics Using SCpubr* Preprint (Bioinformatics, Mar. 2022). doi:10.1101/2022.02.28.482303.

306. Pedersen, T. L. *Patchwork: The Composer of Plots* Manual (2022).

307. Martin Morgan [Cre], M. C. [*AnnotationHub* Bioconductor. 2017. doi:10.18129/B9.BIOC.ANNOTATIONHUB.

308. Stickels, R. R. *et al.* Highly Sensitive Spatial Transcriptomics at Near-Cellular Resolution with Slide-seqV2. *Nature biotechnology* **39,** 313–319. doi:10.1038/s41587-020-0739-1 (Mar. 2021).

309. Dobin, A. *et al.* STAR: Ultrafast Universal RNA-seq Aligner. *Bioinformatics* **29,** 15–21. doi:10.1093/bioinformatics/bts635 (Jan. 2013).

310. Hopfensitz, M., Müssel, C., Maucher, M. & Kestler, H. A. Attractors in Boolean Networks: A Tutorial. *Computational Statistics* **28,** 19–36. doi:10.1007/s00180-012-0324-2 (Feb. 2013).

311. Huang, S. Gene Expression Profiling, Genetic Networks, and Cellular States: An Integrating Concept for Tumorigenesis and Drug Discovery. *Journal of molecular medicine* **77,** 469–480 (1999).

312. Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. The Yeast Cell-Cycle Network Is Robustly Designed. *Proceedings of the National Academy of Sciences* **101,** 4781–4786 (2004).

313. Grieco, L. *et al.* Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision. *PLoS Computational Biology* **9,** e1003286. doi:10.1371/journal.pcbi.1003286 (2013).

314. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP Kinase Signalling Pathways in Cancer. *Oncogene* **26,** 3279–3290. doi:10.1038/sj.onc.1210421 (May 2007).

315. Wagner, E. F. & Nebreda, A. R. Signal Integration by JNK and P38 MAPK Pathways in Cancer Development. *Nature Reviews. Cancer* **9,** 537–549. doi:10.1038/nrc2694 (Aug. 2009).

316. Müssel, C., Hopfensitz, M. & Kestler, H. A. BoolNet—an R Package for Generation, Reconstruction and Analysis of Boolean Networks. *Bioinformatics (Oxford, England)* **26,** 1378–1380. doi:10.1093/bioinformatics/btq124 (May 2010).

317. Biere, A. PicoSAT Essentials. *Journal on Satisfiability, Boolean Modeling and Computation* **4,** 75–97 (2008).

318. Dubrova, E. & Teslenko, M. A SAT-based Algorithm for Finding Attractors in Synchronous Boolean Networks. *IEEE/ACM transactions on computational biology and bioinformatics* **8,** 1393–1399. doi:10.1109/tcbb.2010.20 (Sept. 2011).

319. Heldin, C.-H., Landström, M. & Moustakas, A. Mechanism of TGF-β Signaling to Growth Arrest, Apoptosis, and Epithelial–Mesenchymal Transition. *Current Opinion in Cell Biology* **21,** 166–176. doi:10.1016/j.ceb.2009.01.021 (Apr. 2009).

320. Schramek, H. MAP Kinases: From Intracellular Signals to Physiology and Disease. *Physiology* **17,** 62–67 (2002).

321. Kennedy, N. J. & Davis, R. J. Role of JNK in Tumor Development. *Cell Cycle (Georgetown, Tex.)* **2,** 199–201 (2003 May-Jun).

322. Gallo, C. A., Cecchini, R. L., Carballido, J. A., Micheletto, S. & Ponzoni, I. Discretization of Gene Expression Data Revised. *Briefings in Bioinformatics* **17,** 758–770. doi:10.1093/bib/bbv074 (Sept. 2016).

323. Saadatpour, A. & Albert, R. A Comparative Study of Qualitative and Quantitative Dynamic Models of Biological Regulatory Networks. *EPJ Nonlinear Biomedical Physics* **4,** 5. doi:10.1140/epjnbp/s40366-016-0031-y (Dec. 2016).

324. Alvarez, M. J. *et al.* Functional Characterization of Somatic Mutations in Cancer Using Network-Based Inference of Protein Activity. *Nature Genetics* **48,** 838–847. doi:10.1038/ng.3593 (Aug. 2016).

325. Letort, G. *et al.* PhysiBoSS: A Multi-Scale Agent-Based Modelling Framework Integrating Physical Dimension and Cell Signalling. *Bioinformatics* **35,** 1188–1196. doi:10.1093/bioinformatics/bty766 (Apr. 2019).

326. Dagogo-Jack, I. & Shaw, A. T. Tumour Heterogeneity and Resistance to Cancer Therapies. *Nature Reviews Clinical Oncology* **15,** 81–94. doi:10.1038/nrclinonc.2017.166 (Feb. 2018).

327. Victori, P. *Ingres: Inferring Probabilistic Boolean Networks of Gene Regulation Using Protein Activity Enrichment Scores* 2021.

328. Basso, K. *et al.* Reverse Engineering of Regulatory Networks in Human B Cells. *Nature Genetics* **37,** 382–390. doi:10.1038/ng1532 (Apr. 2005).

329. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: Gene Network Reverse Engineering through Adaptive Partitioning Inference of Mutual Information. *Bioinformatics (Oxford, England)* **32,** 2233–2235. doi:10.1093/bioinformatics/btw216 (July 2016).

330. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and Integration of Resources for the Estimation of Human Transcription Factor Activities. *Genome Research* **29,** 1363–1375. doi:10.1101/gr.240663.118 (Aug. 2019).

331. Ding, H. *et al.* Quantitative Assessment of Protein Activity in Orphan Tissues and Single Cells Using the metaVIPER Algorithm. *Nature Communications* **9,** 1471. doi:10.1038/s41467-018-03843-3 (Apr. 2018).

332. Giorgi, F. M. *Aracne.Networks: ARACNe-inferred Gene Networks from TCGA Tumor Datasets* Bioconductor. 2017. doi:10.18129/B9.BIOC.ARACNE.NETWORKS.

333. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21. doi:10.1016/j.cell.2019.05.031 (June 2019).

334. Petersen, T. L. *Tidygraph: A Tidy API for Graph Manipulation* 2020.

335. Chaouiya, C., Naldi, A. & Thieffry, D. Logical Modelling of Gene Regulatory Networks with GINsim. *Methods in Molecular Biology* **804,** 463–479. doi:10.1007/978-1-61779-361-5_23 (2012).

336. Remy, E. *et al.* A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. *Cancer Research* **75,** 4042–4052. doi:10.1158/0008-5472.CAN-15-0602 (Oct. 2015).

337. *Cell-Cycle Scoring and Regression*

338. Wang, L. *et al.* A Reference Profile-Free Deconvolution Method to Infer Cancer Cell-Intrinsic Subtypes and Tumor-Type-Specific Stromal Profiles. *Genome Medicine* **12,** 24. doi:10.1186/s13073-020-0720-0 (Dec. 2020).

339. Jiang, Y., Pjesivac-Grbovic, J., Cantrell, C. & Freyer, J. P. A Multiscale Model for Avascular Tumor Growth. *Biophysical Journal* **89,** 3884–3894. doi:10.1529/biophysj.105.060640 (Dec. 2005).

340. Liu, Q. *et al.* EGFR-TKIs Resistance via EGFR-independent Signaling Pathways. *Molecular Cancer* **17,** 53. doi:10.1186/s12943-018-0793-1 (Dec. 2018).

341. Bertotti, A. & Sassi, F. Molecular Pathways: Sensitivity and Resistance to Anti-EGFR Antibodies. *Clinical Cancer Research* **21,** 3377–3383. doi:10.1158/1078-0432.CCR-14-0848 (Aug. 2015).

342. Guerrab, A. E. *et al.* Anti-EGFR Monoclonal Antibodies and EGFR Tyrosine Kinase Inhibitors as Combination Therapy for Triple-Negative Breast Cancer. *Oncotarget* **7.** doi:10.18632/oncotarget.12037 (Nov. 2016).

343. Mueller, K. L., Yang, Z.-Q., Haddad, R., Ethier, S. P. & Boerner, J. L. EGFR/Met Association Regulates EGFR TKI Resistance in Breast Cancer. *Journal of Molecular Signaling* **5,** 8. doi:10.1186/1750-2187-5-8 (July 2010).

344. Karamouzis, M. V., Konstantinopoulos, P. A. & Papavassiliou, A. G. Targeting MET as a Strategy to Overcome Crosstalk-Related Resistance to EGFR Inhibitors. *The Lancet Oncology* **10,** 709–717. doi:10.1016/S1470-2045(09)70137-8 (July 2009).

345. Oliveira, M. *et al.* FAIRLANE, a Double-Blind Placebo-Controlled Randomized Phase II Trial of Neoadjuvant Ipatasertib plus Paclitaxel for Early Triple-Negative Breast Cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology* **30,** 1289–1297. doi:10.1093/annonc/mdz177 (Aug. 2019).

346. Costa, R. L. B., Han, H. S. & Gradishar, W. J. Targeting the PI3K/AKT/mTOR Pathway in Triple-Negative Breast Cancer: A Review. *Breast Cancer Research and Treatment* **169,** 397–406. doi:10.1007/s10549-018-4697-y (June 2018).

347. Khan, M. A., Jain, V. K., Rizwanullah, M., Ahmad, J. & Jain, K. PI3K/AKT/mTOR Pathway Inhibitors in Triple-Negative Breast Cancer: A Review on Drug Discovery and Future Challenges. *Drug Discovery Today* **24,** 2181–2191. doi:10.1016/j.drudis.2019.09.001 (Nov. 2019).

348. Xiao, X., Li, B., Mitton, B., Ikeda, A. & Sakamoto, K. Targeting CREB for Cancer Therapy: Friend or Foe. *Current Cancer Drug Targets* **10,** 384–391. doi:10.2174/156800910791208535 (June 2010).

349. Ramkissoon, A. *et al.* Targeted Inhibition of the Dual Specificity Phosphatases DUSP1 and DUSP6 Suppress MPNST Growth via JNK. *Clinical Cancer Research* **25,** 4117–4127. doi:10.1158/1078-0432.CCR-18-3224 (July 2019).

350. Wu, R. *et al.* Aurovertin B Exerts Potent Antitumor Activity against Triple-Negative Breast Cancer in Vivo and in Vitro via Regulating ATP Synthase Activity and DUSP1 Expression. *Die Pharmazie - An International Journal of Pharmaceutical Sciences* **75,** 261–265. doi:10.1691/ph.2020.0380 (June 2020).

351. Hauge, C. & Frödin, M. RSK and MSK in MAP Kinase Signalling. *Journal of Cell Science* **119,** 3021–3023. doi:10.1242/jcs.02950 (Aug. 2006).

352. Huang, S. *et al.* Exosomal miR-130b-3p Targets SIK1 to Inhibit Medulloblastoma Tumorigenesis. *Cell Death & Disease* **11,** 1–16. doi:10.1038/s41419-020-2621-y (June 2020).

353. Fan, Y. *et al.* Dual-Target MDM2/MDMX Inhibitor Increases the Sensitization of Doxorubicin and Inhibits Migration and Invasion Abilities of Triple-Negative Breast Cancer Cells through Activation of TAB1/TAK1/P38 MAPK Pathway. *Cancer Biology & Therapy* **20,** 617–632. doi:10.1080/15384047.2018.1539290 (May 2019).

354. Bai, R. *et al.* MicroRNA-409-3p Suppresses Colorectal Cancer Invasion and Metastasis Partly by Targeting GAB1 Expression. *International Journal of Cancer* **137,** 2310–2322. doi:10.1002/ijc.29607 (2015).

355. Wang, E. *et al.* Tumor Penetrating Peptides Inhibiting MYC as a Potent Targeted Therapeutic Strategy for Triple-Negative Breast Cancers. *Oncogene* **38,** 140–150. doi:10.1038/s41388-018-0421-y (Jan. 2019).

356. Wang, N. *et al.* The Landscape of PDK1 in Breast Cancer. *Cancers* **14,** 811. doi:10.3390/cancers14030811 (Jan. 2022).

357. Hollestelle, A. *et al.* Distinct Gene Mutation Profiles among Luminal-Type and Basal-Type Breast Cancer Cell Lines. *Breast Cancer Research and Treatment* **121,** 53–64. doi:10.1007/s10549-009-0460-8 (May 2010).

358. Paterson, Y. Z. *et al.* A Toolbox for Discrete Modelling of Cell Signalling Dynamics. *Integrative Biology* **10,** 370–382. doi:10.1039/C8IB00026C (2018).