

### Revealing Putative Drug Targets for Basal-Like Breast Cancer

Sharon Nienyun Hsu

Submitted for the degree of

Doctor of Philosophy at the University of Leeds January 2021

### Summary

The development of targeted drugs has revolutionised the treatment strategy for breast cancer, improving patients' clinical outcome and quality of life. However, the lack of targetable proteins has limited treatment options for patients with Basal-like breast cancer to non-selective cytotoxic and cytostatic drugs. This research aimed to reveal putative drug targets for Basal-like breast cancer.

The advances of sequencing technologies and computational methods have benefited drug target identification research, which is a critical early step in the drug discovery process. The method developed in this research integrated omic data and modelling approaches to uncover the transcriptional characteristics underlying Basal-like breast cancer. This molecular characterisation revealed potential candidates for further pharmaceutical intervention.

To reveal putative drug targets for Basal-like breast cancer, an unsupervised clustering approach was first performed to study the heterogeneity in breast cancer. The clustering analysis revealed several differently expressed transcriptional factors in Basal-like breast cancer. Using a network modelling approach, these transcriptional factors were then prioritised according to their topological features. Assuming gene expression is the first proxy of protein expression, the expression correlations between critical candidate genes were explored at the protein level. Using bioinformatics platforms and databases, several putative protein-protein interactions that are associate with candidate genes were identified. This was followed by a specific target protein selection. Furthermore, a molecular docking analysis was performed to determine the structural characteristics and molecular recognition features of these interactions. Finally, a laboratory experiment was developed to evaluate the protein interactions in breast cancer cell lines.

This research has not only identified putative drug targets for Basal-like, but also developed an integrative target identification approach that can be adapted easily to study other diseases.

## Preface

This dissertation is the result of my own research in the School of Food Science and Nutrition and Astbury Centre for Structural Molecular Biology, University of Leeds, from October 2016 to November 2020. All the work presented in this thesis is original, unless otherwise stated. None of the work has been submitted for another degree, diploma or other qualification at this or any other University.

© 2021 The University of Leeds, Sharon Nienyun Hsu

Signed

### Acknowledgements

I am deeply grateful to my supervisor, Dr James Smith, for his guidance, advice and support throughout this project. He has always been there when I needed support and his encouragement has meant a lot to me, especially at the beginning of my PhD. I feel privileged and honoured to have the chance to work with him and his enthusiasm in science and research has inspired and motivated me.

My appreciation also goes to my co-supervisor Professor Richard Bayliss, who has provided insightful discussions and suggestions. I am also grateful to Dr Josephina Sampson, for her patient and constant support in guiding me to do the laboratory experiments. I would also like to thank the project students, especially Yaowei Xun, for contributing to this research. Their effort and input were very appreciated. My gratitude extends to all, past and present, members of Smith's and Bayliss' group, for their helpful contributions and making this journey special and enjoyable.

I would also like to thank Professor Valerie Spiers for her valuable discussions and kindness for providing me breast cancer cell lines. I am also grateful to Professor Chi-Ying Huang for his useful discussion and advice on possible extension of this work. I would also like to thank Dr Helen Chappell who has inspired and encouraged me throughout the frustrating times.

I would like to thank all of my friends in Leeds, London, Taiwan and other part of the world for their support and this especially goes to Andy, Christine, Alessandra, Wendy, Yahua, Aiai and Manman. A special thanks to Tom for his unwavering support and believing in me throughout this journey. Last but not least, I am indebted to my Mum, Dad, Nick and grandparents for supporting, encouraging and believing in me throughout these years and made this journey possible.

## Contents

$\mathbf{A}$	Abbreviations xv			
1	Intr	roduction		
	1.1	Breast	cancer	2
		1.1.1	Epidemiology	2
		1.1.2	Risk factors	2
		1.1.3	Prognosis	3
		1.1.4	Molecular classification of breast cancer	5
		1.1.5	Basal-like breast cancer	9
		1.1.6	Treatments	11
	1.2	Nuclea	ar receptors	14
		1.2.1	Nuclear receptor superfamily	14
		1.2.2	Coregulators	19
		1.2.3	Involvement in breast cancer	21
	1.3	Comp	uter-aided drug discovery	23
		1.3.1	Drug discovery process	24
		1.3.2	Bioinformatics in drug target identification	26
	1.4	Chapt	er summary	29
<b>2</b>	Bre	ast Ca	ncer Subtype Clustering	<b>31</b>
	2.1	Introd	$uction \ldots \ldots$	31
		2.1.1	Patient clustering in cancer studies	31
		2.1.2	Model-based clustering vs classical clustering	32
		2.1.3	Network modelling	34
		2.1.4	Aims	35

		2.1.5	Chapter overview	36
	2.2	Cluste	ring materials and methods	37
		2.2.1	TCGA expression data	37
		2.2.2	METABRIC expression data	38
		2.2.3	Unsupervised Bayesian Hierarchical Clustering	39
		2.2.4	Partial correlation networks	40
		2.2.5	Hub-associated local networks and 3-node motifs	41
	2.3	Strati	fication comparison by BHC	41
		2.3.1	Basal-like vs Luminal A	42
		2.3.2	Basal-like vs Luminal B	44
		2.3.3	Basal-like vs Her2	46
	2.4	Basal-	specific correlation networks	48
	2.5	Motif	classifications and identification	49
	2.6	Discus	ssions and summary	59
		2.6.1	Heterogeneity in breast cancer subtypes	60
		2.6.2	Partial correlation-based network and choice of thresholding	61
		2.6.3	Implication of 3-node motifs	63
		2.6.4	Limitations and future perspectives	66
	2.7	Key fi	ndings	67
3	Ger	ne Exp	ression Clustering	69
	3.1	Introd	uction	69
		3.1.1	Targeting Basal-like subtype	69
		3.1.2	Aims	71
		3.1.3	Chapter overview	71
	3.2	Metho	ds	72
		3.2.1	BHC clustering and identification of Basal Candidate Genes	72
		3.2.2	Basal-specific correlation network	73
	3.3	Gene	clustering results	73
		3.3.1	Drug target assessment	74
		3.3.2	Clustering results	76
	3.4	Basal	candidate gene identification	78

	3.5	Basal	associated correlation networks	79
	3.6	Discus	sion and future perspectives	82
	3.7	Key fi	m ndings	84
4	Ider	ntificat	ion of Critical Protein-Protein Interactions	85
	4.1	Introd	uction	85
		4.1.1	Protein-protein interactions	85
		4.1.2	Targeting protein-protein interactions	86
		4.1.3	Aims	87
		4.1.4	Chapter overview	87
	4.2	Metho	ds and databases	88
		4.2.1	Protein-protein interaction networks	88
		4.2.2	Protein structural information and SLiMs identification	89
	4.3	LCK-S	STAT1 correlation	89
		4.3.1	Approach 1: LCK-x-y-STAT1	89
		4.3.2	Approach 2: LCK and STAT1 PPI network	93
	4.4	Putati	ve binding regions	94
		4.4.1	KIT in breast cancer	94
		4.4.2	GAB1 in breast cancer	96
		4.4.3	Putative binding sites	98
		4.4.4	Prediction of structural domains in LCK and GAB1's binding partners .	101
		4.4.5	SH2 domain alignment	102
		4.4.6	Structural prediction of KIT and GAB1	103
		4.4.7	SLiMs implicated in IDRs	104
	4.5	Discus	sion	106
		4.5.1	Gene-protein correlation	106
		4.5.2	Reliability of protein-protein interactions and target validation	108
		4.5.3	Proteins with tandem SH2 domains	110
		4.5.4	Tandem SH2 protein binding models	110
	4.6	Key fi	m ndings	114
5	Stru	uctural	Characterisation	115
	5.1	Introd	uction	115

6

	5.1.1	Diversity of protein-protein interactions
	5.1.2	Intrinsically disorder regions in signalling proteins
	5.1.3	SH2 domains
	5.1.4	Aims
	5.1.5	Overview
5.2	Data	collection and docking protocols
	5.2.1	Available PDB crystal structures
	5.2.2	Data preparation
	5.2.3	Preliminary stage
	5.2.4	Peptide docking
	5.2.5	Conformational energy distribution
	5.2.6	Focussed density contour plots
	5.2.7	Representative structure evaluation
5.3	An en	semble of KIT-PI3K interactions
5.4	An en	semble of GAB1-PI3K interactions
	5.4.1	GAB1 444-450 to PI3K regulatory subunits
	5.4.2	GAB1 469-475 to PI3K regulatory subunits
	5.4.3	GAB1 586 to 592 to PI3K regulatory subunits
5.5	Molec	ular recognition of SH2 domains in PI3K
5.6	Discus	ssion $\ldots \ldots \ldots$
	5.6.1	Biological roles
	5.6.2	Role of water in protein interactions
	5.6.3	Exploring flanking regions
	5.6.4	Future challenges and perspectives
5.7	Key fi	ndings
Exp	oerime	ntal Validation 162
6.1	Introd	luction $\ldots \ldots \ldots$
	6.1.1	Protein kinases in cancer
	6.1.2	KIT
	6.1.3	PI3K regulatory subunits
	6.1.4	Aims

		6.1.5	Chapter overview	. 166
	6.2	Mater	ials and methods	. 167
		6.2.1	Cell culture	. 167
		6.2.2	Cell passage	. 168
		6.2.3	Cell freezing	. 169
		6.2.4	Thawing cells	. 169
		6.2.5	Cell seeding and fixing	. 169
		6.2.6	Indirect dual immunofluorescence protocol	. 169
		6.2.7	In situ Proximity Ligation Assay protocol	. 171
		6.2.8	Antibody combinations	. 172
	6.3	Result	js	. 174
		6.3.1	Protein sub-cellular localisation	. 174
		6.3.2	In Situ Proximity Ligation Assay results	. 184
	6.4	Discus	ssion and speculative research	. 187
		6.4.1	Cellular localisation of KIT, PIK3R1 and PIK3R2	. 187
		6.4.2	KIT and PIK3R1/2 protein interactions and binding sites $\ldots$ $\ldots$ $\ldots$	. 190
		6.4.3	Speculative research and future work	. 195
	6.5	Key fi	$\operatorname{ndings}$	. 198
7	Disc	cussion	and Future work	199
Re	efere	nces		205
Aj	ppen	dix A	Bayesian Hierarchical Clustering Algorithm	267
Aj	ppen	dix B	Nuclear Receptors and Coregulators	<b>2</b> 81
Aj	ppen	dix C	Nuclear Receptor Availability	290
Aj	ppen	dix D	Patient Clustering Consolidation	297
Aj	ppen	dix E	Partial Correlation Networks	299
Aj	ppen	dix F	Nodes and Degree	307
A	ppen	dix G	Networks with Different Thresholdings	311

Appendix H	Highly Expressed G	enes in Basal-like	Breast Cancer	318
Appendix I	Docking Flags		:	321

# List of Figures

1.1	Venn diagram representing the relationship between triple negative breast cancer	
	and Basal-like breast cancer.	9
1.2	NR-mediated signalling pathways that have been implicated in breast cancer $\ . \ .$	22
1.3	Stages of the drug discovery process	24
2.1	Overview of the approach presented in this chapter.	36
2.2	Schematic illustration of subtype dominant class and ambiguous class	40
2.3	BHC results from Basal-like vs Luminal A comparison	43
2.4	BHC results from Basal-like vs Luminal B comparison	45
2.5	BHC results from Basal-like vs Her2 comparison	47
2.6	An example of Basal-specific partial correlation networks	49
2.7	Node degree summary	51
2.8	FOS-centred local networks from TCGA analyses.	52
2.9	FOS-centred local networks from METABRIC analyses	52
2.10	STAT1-centred local networks from TCGA analyses	53
2.11	STAT1-centred local networks from METABRIC analyses	53
2.12	NPU classification of 3-node motifs.	55
2.13	Linear and complete motifs	56
2.14	Motif JUN-FOS-NR4A2 embedded in the six out of eight Basal-specific networks.	58
2.15	Motif PML-STAT1-GCH1 embedded in the six out of eight Basal-specific networks.	58
2.16	${\it Motif STAT1-LCK-NR1H3}\ embedded\ in\ the\ six\ out\ of\ eight\ Basal-specific\ networks.}$	59
2.17	Comparison between canonical correlation network and partial correlation network.	62
2.18	A correlation network motif and its possible regulatory mechanisms	67
3.1	Overview of the chapter.	72

3.2	Drug target containing gene clusters revealed by BHC
3.3	Gene clusters in Basal-like subtype
3.4	Venn diagram of the three sets of highly expressed genes extracted from the
	paired clustering analyses
3.5	Partial correlation network derived from Basal-like patients
4.1	Chapter overview
4.2	Illustration of LCK and STAT1 PPI networks
4.3	KIT signalling pathways
4.4	GAB1 signalling pathways
4.5	Schematic representation of KIT and GAB1's binding partners
4.6	Sequence alignment of N-SH2 and C-SH2 domains
4.7	Predicted structural domains in KIT and GAB1
4.8	The predicted SLiMs from the short peptide sequences in KIT and GAB1 105
4.9	A schematic illustrating two proposed protein-protein interaction models of KIT,
	GAB1 and PI3K subunits
4.10	Schematic illustration of a three-state equilibrium SH2 binding model in PI3K 113 $$
5.1	Peptide docking workflow
5.2	Process for identifying relevant pre-solved crystal structures from the PDB 121
5.3	A list of pY peptides and SH2 domains considered in Chapter 5
5.4	Scatter and density contour plots from KIT 718-725 to PI3K SH2 domain simu-
	lations
5.5	Representative structures revealed from KIT 718-725 to SH2 domains in PIK3R1.129
5.6	Representative structures revealed from KIT 718-725 to SH2 domains in PIK3R2.131
5.7	Representative structures revealed from KIT 718-725 to SH2 domains in PIK3R3.133
5.8	Scatter and density contour plots from GAB1 444-450 to PI3K SH2 domain
	simulations
5.9	Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R1.137
5.10	Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R2.139 $$
5.11	Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R3.141
5.12	Scatter and density contour plots from GAB1 469-475 to PI3K SH2 domain
	simulations

5.13	Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R1.145 $$
5.14	Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R2.146 $$
5.15	Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R3.148 $$
5.16	Scatter and density contour plots from GAB1 586-592 to PI3K SH2 domain
	simulations
5.17	Representative structures revealed from GAB1 586-592 to SH2 domains in PIK3R1 $$
	and PIK3R2
5.18	Representative structures revealed from GAB1 586-592 to SH2 domains in PIK3R3.154 $$
5.19	Characterisation of pY peptide binding modes
5.20	Beta-C strands in SH2 domains in PI3K regulatory subunits
6.1	Schematic illustration of Class IA PI3K activation
6.2	Chapter overview
6.3	Schematic illusion of antibody pairs examined in this chapter
6.4	Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in MDA-MB-231.176 $$
6.5	Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in MDA-MB-468.179 $$
6.6	Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in LGI1T 180 $$
6.7	Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in HBL100. $\ .$ . 182
6.8	PLA analysis for the four antibody pairs in four cell lines
6.9	PLA signal scatter plots
6.10	Schematic illustration the concept of antibody binding and epitopes exposure in
	different form of protein-protein interactions
6.11	Schematic diagram of proposed work
A.1	Schematic of a clustering tree
A.2	Bayesian hierarchical clustering algorithm
D.1	Clusters to classes simplification from the Basal vs Luminal A analyses 297
D.2	Clusters to classes simplification from the Basal vs Luminal B analyses 297
D.3	Clusters to classes simplification from the Basal vs Her2 analyses
E.1	The partial correlation network derived from TCGA class 1 from the Basal vs
	Luminal A analysis.

E.2	The partial correlation network derived from METABRIC class 1 from the Basal
	vs Luminal A analysis
E.3	The partial correlation network derived from TCGA class 1 from the Basal vs
	Luminal B analysis
E.4	The partial correlation network derived from METABRIC class 1 from the Basal
	vs Luminal B analysis
E.5	The partial correlation network derived from METABRIC class 2 from the Basal
	vs Luminal B analysis
E.6	The partial correlation network derived from TCGA class 1 from the Basal vs
	Her2 analysis
E.7	The partial correlation network derived from TCGA class 3 from the Basal vs
	Her2 analysis
E.8	The partial correlation network derived from METABRIC class 2 from the Basal
	vs Her2 analysis
G.1	The partial correlation network constructed using a $0.5\%$ thresholding, containing
	72 edges
G.2	The partial correlation network constructed using a $1\%$ thresholding, containing
	145 edges
G.3	The partial correlation network constructed using a $2\%$ thresholding, containing
	291 edges
G.4	The partial correlation network constructed using a $5\%$ thresholding, containing
	727 edges
G.5	The partial correlation network constructed using a $10\%$ thresholding, containing
	1453 edges
G.6	The partial correlation network constructed using a $15\%$ thresholding, containing
	2180 edges

## List of Tables

1.1	Summary of the breast cancer molecular subtypes
1.2	Nuclear receptor families
1.3	Databases and used in this project
2.1	Sample subtype distribution
2.2	Occurrence of the seven types of NPU motifs centred around FOS and STAT1
	from Basal-specific networks
2.3	Occurrence of the critical 3-node motifs enriched in Basal-specific networks 59
3.1	Degree of each Basal candidate genes in the basal partial correlation network 80
4.1	Proteins that satisfy the LCK-x-y-STAT1 configuration
4.2	Details of the 14 LCK-x-y-STAT1 PPI configurations
4.3	Summary of previously reported binding sites for KIT and GAB1 PPIs 100
4.4	N-SH2 and C-SH2 domains in KIT and GAB1's binding partners
5.1	Details of the relevant crystal structures collected from PDB
5.2	The values for 90% of the x and y distributions for each KIT 718-725 to PI3K
	SH2 domain simulation
5.3	The values for $90\%$ of the x and y distributions for each GAB1 444-450 to PI3K
	SH2 domain simulation
5.4	The values for $90\%$ of the x and y distributions for each GAB1 478-484 to PI3K
	SH2 domain simulation
5.5	The values for $90\%$ of the x and y distributions for each GAB1 586-592 to PI3K
	SH2 domain simulation
6.1	Details of the four cell lines used in this chapter

6.2	Details of antibodies used in this chapter
6.3	Summary of protein localisations
B.1	The 48 human nuclear receptors, with their corresponding ligands and coregu-
	lators. There are a total number of 178 non-redundant genes, including the $48$
	nuclear receptors and coregulators
H.1	Genes highly expressed in Basal-like breast cancer when compared to Luminal A. 318
H.2	Genes highly expressed in Basal-like breast cancer when compared to Luminal B. 319
H.3	Genes highly expressed in Basal-like breast cancer when compared to Her2 320

## List of Abbreviations

AA amino acid
<b>AKT</b> protein kinase B
<b>ALK</b> anaplastic lymphoma kinase
AP-1 activator protein 1
$\mathbf{AR}$ and rogen receptor
ATG8 autophagy-related protein 8
<b>ATM</b> ATM serine/threenine kinase
Auroa aurora kinase
<b>BCGs</b> basal candidate genes
<b>BCL11A</b> BAF chromatin remodeling complex subunit BCL11A
<b>BCR-ABL</b> breakpoint cluster region and tyrosine-protein kinase ABL1 fusion protein
<b>BHC</b> bayesian hierarchical clustering
<b>BLBC</b> basal-like breast cancer
<b>BPE</b> bovine pituitary extract
${\bf BRAF}$ serine/threonine-protein kinase B-Raf
BRCA1 breast cancer type 1 protein
<b>BRCA2</b> breast cancer type 2 protein

 ${\bf BSA}\,$  bovine serum albumin

CDH1 Cadherin-1

CHECK2 checkpoint kinase 2

 ${\bf C}{\bf K}$  creatine kinase

 $\mathbf{CK14}$  creatine kinase 14

 ${\bf CK17}\,$  creatine kinase 17  $\,$ 

 ${\bf CK5}\,$  creatine kinase 5  $\,$ 

CK6 creatine kinase 6

DAPI 4, 6-diamidino-2-phenylindole

 ${\bf DMEM}\,$  culture media dulbecco?s modified eagle medium

 ${\bf DMSO}\,$  dimethyl sulfoxide

 ${\bf EGF}$  epidermal growth factor

 ${\bf EGFR}\,$  epidermal growth factor receptor

**ER** oestrogen receptor (note: US spelling 'estrogen receptor')

**ER1** oestrogen receptor 1

**ER2** oestrogen receptor 2

**ERBB2** erb-B2 Receptor Tyrosine Kinase 2

 ${\bf ERK}$  extracellular signal-regulated kinases

ERRs oestrogen receptor-related receptors

ESRRA oestrogen related receptor alpha

FABP6 fatty acid binding protein 6

 ${\bf FBS}\,$  fetal bovine serum

 ${\bf FDA}\,$  food and drug administration

 ${\bf FHA}$  for khead-associated domain

FOS fos proto-Oncogene

GAB1 GRB2 associated binding Protein 1 **GAPs** GTPase activating proteins GC glucocorticoids GCH1 GTP cyclohydrolase 1 **GCNs** gene correlation networks GPCRs G-protein-coupled receptors **GR** glucocorticoid receptor Her2 human epidermal growth factor receptor 2 **HGF** hepatocyte growth factor HSP90AB1 heat shock protein 90 alpha family class B member 1 **IAP** inhibitors of apoptosis proteins  $IC_{50}$  half-maximal inhibitory concentration **IDRs** intrinsically disordered regions **IEGs** immediate early genes **IF** immunofluorescence **IFN** interferon IMP3 insulin-like growth factor II mRNA binding protein 3 JAK janus kinase JUN jun Proto-Oncogene KHDRBS1 KH RNA binding domain containing signal transduction associated 1 **KIT** stem cell growth factor receptor Kit **LAT** linker for activation of T Cells LCK lymphocyte-specific protein tyrosine kinase

 $\mathbf{MAPK}\xspace$  mitogen-activated protein kinase

 $\mathbf{MC}$  mineralocorticoidss

 $\mathbf{MET}$  tyrosine-protein kinase MET

METABRIC molecular taxonomy of breast cancer international consortium

 $\mathbf{mTOR}\ \mathrm{mechanistic}\ \mathrm{target}\ \mathrm{of}\ \mathrm{rapamycin}$ 

NGS next-generation sequencing

NPAS2 neuronal PAS domain protein 2

**NPI** nottingham prognostic index

 ${\bf NR}\,$  nuclear receptor

NR1I2 nuclear receptor subfamily 1 group I member 2

NR2C2 nuclear receptor subfamily 2 group C member 2

 ${\bf NR2E1}$  nuclear receptor TLX

NR3C1 nuclear receptor subfamily 3 group C member 1

 $\mathbf{NR3C2} \ \mathrm{mineralocorticoid} \ \mathrm{receptor}$ 

NRF1 nuclear respiratory factor 1

 ${\bf NRs}\,$  nuclear receptors

NURR1 nuclear receptor related 1 protein

**p53** tumor protein P53

PALB2 partner and localizer of BRCA2

**PARP** Poly ADP-ribose polymerase

PAX6 paired box protein Pax-6

 ${\bf PBS}\,$  phosphate-buffered saline

**PDB** protein data bank

**PDGF** platelet-derived growth factor receptor

**PH** pleckstrin homology

**PI3K** phosphatidylinositol 3-kinase

PIK3R1 PI3K regulatory subunit 1

PIK3R2 PI3K regulatory subunit 2

PIK3R3 PI3K regulatory subunit 3

**PIP2** phosphorylate phosphatidylinositol 3, 4-bis-phosphate

**PIP3** phosphatidylinositol 3,4,5 trisphosphate

pKIT phosphorylated KIT receptor tyrosine kinase

 $\mathbf{PLA}\ \mathrm{proximity}\ \mathrm{ligation}\ \mathrm{assay}$ 

PLCG1 phospholipase C Gamma 1

**PML** promyelocytic leukemia protein

 $\mathbf{POMC}$  pro-opiomelanocortin

**PPARA** peroxisome proliferator-activated receptor alpha

**PPARGC1A** peroxisome proliferator-activated receptor gamma coactivator 1-alpha

**PPARGC1B** peroxisome proliferator-activated receptor gamma coactivator 1-beta

**PPARs** peroxisome proliferator-activated receptors

**PPI** protein-protein interaction

**PPIs** protein-protein interaction networks

**PR** progesterone receptor

**PTEN** phosphatase and tensin homolog

**PTPN11** protein tyrosine phosphatase non-receptor type 11

**PTPN22** protein tyrosine phosphatase non-receptor type 22

**PTPRC** protein tyrosine phosphatase receptor type C

 $\mathbf{p}\mathbf{Y}$  phosphorylated tyrosine

 $\mathbf{RAR}$  retinoid receptor

**RARB** retinoic Acid Receptor Beta  ${\bf RARs}\,$  retinoid receptors Ras Ras GTPase **RASA1** RAS P21 protein activator 1 **RMSD** root mean squared deviation **RPMI 1640** roswell park memorial institute medium **RTKs** receptor tyrosine kinases **RXR** retinoid X receptor  $\mathbf{SCF}$  stem cell factor **SERM** selective oestrogen receptor modulators SH2 Src Homology 2 SH2D2A SH2 domain containing 2A SH3 Src homology 3 Shp2 protein tyrosine phosphatase SIX3 homeobox protein SIX3 **SLiMs** short linear motifs SOX9 SRY-Box transcription factor 9 **STAR** steroidogenic acute regulatory protein **STAT** transducer and activator of transcription **STAT1** signal transducer and activator of transcription 1 **STAT3** signal transducer and activator of transcription 3 **SYK** spleen associated tyrosine kinase TAF15 TATA-binding protein-associated factor 2N

**TCF3** transcription factor 3

**TCGA** the cancer genome atlas

 $\mathbf{TCR}~$  T-cell antigen receptor

**THR** thyroid hormone receptor

**TMPRSS2** transmembrane protease serine 2

**TNBC** triple negative breast cancer

**TP53** tumor protein P53

 ${\bf TPX2}~{\rm TPX2}$  microtubule nucleation factor

VDR vitamin D3 receptor

 ${\bf VEGF}$  vascular endothelial growth factor

 ${\bf VEGFR}\,$  vascular endothelial growth factor receptor

 $\mathbf{WT1}$  wilms tumour 1

ZAP70 zeta chain of T cell receptor associated protein kinase 70

### Chapter 1

### Introduction

The rapid development of sequencing technologies has allowed comprehensive analyses of complex biological systems. Following the completion of the Human Genome Project in 2003, omics data has been progressively used to better understand the molecular basis of oncogenesis [1]. This new omics era in cancer research is continuing to change the way cancers are predicted, diagnosed and treated. In particular, in the context of cancer treatment, strategies are now moving away from the traditional 'one size fit all' approach to more tailored treatments based on individual's genotype or phynotype[2]. Such an approach is expected to develop and provide more effective and efficient treatment options with fewer side effects, compared to the traditional non-selective treatments. However, such treatment options are not available for all cancers, and unfortunately, Basal-like breast cancer is one of them [3]. The lack of efficient treatment options for Basal-like breast cancer has led to a devastating clinical outcome, affecting a large number of women in the UK and worldwide.

Using multi-omics data and an integrative approach, this thesis describes work that has uncovered the transcriptional activities, in particular from nuclear receptors, underlying Basal-like breast cancer. This molecular characterisation of Basal-like breast cancer has led to the identification of a number of putative drug targets for further pharmaceutical intervention.

In this chapter, the current status, biological and molecular basis of breast cancer will be introduced, with an emphasis on Basal-like subtype. This will be followed by an overview of the roles and involvement of nuclear receptors, a large subset of transcriptional factors, in breast oncogenesis. Finally, I will highlight and discuss how the advances of omics data and computational methods have contributed and played a role in drug target discovery research.

### 1.1 Breast cancer

Breast cancer is a disease in which malignant (cancer) cells form in the breast tissues. Breast cancer cells often developed into a tumour (cluster of malignant cells) which can be felt as a lump or be detected by X-ray scans. Breast cancer is a heterogeneous disease with a high degree of diversity within and between tumours [4]. While it has been demonstrated that breast cancer is heterogeneous at multiple levels, to identify potential molecular drug targets the emphasis of this thesis has been to explore the molecular heterogeneity underlying breast cancer.

#### 1.1.1 Epidemiology

Breast cancer accounts for 11.6% of new cancer cases worldwide, making it the most commonly diagnosed cancer in women [5]. In the UK, there are approximately 55,200 new breast cancer cases each year (2015-2017 statistics), affecting 1 in 7 women in their lifetime. While it mostly affects women, it can also affect men with around 390 new male breast cancer cases per year. Breast cancer is now recognised as one of the most treatable cancers, with the mortality rates decreased by almost 40% since the 1970s and 20% over the last decade. Indeed the breast cancer survival rate in the UK has doubled in the last 40 years. The latest estimation (2017) suggests that more than three-quarters of women survive their disease for ten years or more after diagnosis. However, there are still around 11,400 breast cancer incident rate has increased both in the UK and worldwide. It has been reported that in the UK, the incidence rate increased by 18% between 1933-1996 and 2014-2017. The prevalence of this cancer has made it a popular subject of research. The UK statistics were obtained and summarised from Cancer Research UK, and the National Cancer Registration and Analysis Service (Public Health England) [6, 7].

#### 1.1.2 Risk factors

Apart from being a woman, other risk factors associated with breast cancer include age, obesity, genetic, familial history, lifestyle (i.e. high alcohol consumption, lack of exercise), radiation exposure to the chest before age 30, pregnancy history, breastfeeding history and menstrual history [6]. About 5% to 10% of breast cancer are thought to be hereditary, which can be

caused by abnormal genetic alterations passed from parent to child [8]. The most common (about 10%) inherited mutation in breast cancers are breast cancer type 1 protein (BRCA1) and breast cancer type 2 protein (BRCA2) mutations that increase the lifetime risk of developing the cancer by up to 90% [9]. Other inherited gene mutations have also been linked to breast cancer, including mutations in tumor protein P53 (p53) [10], phosphatase and tensin homolog (PTEN) [11] and partner and localizer of BRCA2 (PALB2) [12] gene. While these mutations are less common than BRCA genes, they all have been shown to be associated with an increased breast cancer risk. Other rarer genetic variants that influence breast cancer prevalence include ATM serine/threenine kinase (ATM) [13], Cadherin-1 (CDH1)[14] and checkpoint kinase 2 (CHECK2) [15]. Family history of breast cancer is also a risk factor. A woman's risk of developing breast cancer is increased if one of her first-degree female relatives is diagnosed with breast cancer, and this risk doubled when more close relatives have breast cancer [16]. Women with a strong family history or have one or more pathogenic mutations may benefit from risk managing precautions, such as regular screening, prophylactic treatments (i.e. tamoxifen and raloxifene), mastectomy surgery and lifestyle changes. With 37.7% of cancer cases in the UK attributed to modifiable risk factors, lifestyle factors are also thought to contribute substantially to the risk of developing breast cancer [17]. These factors are often associated with personal behaviours and can be preventable through lifestyle changes. For example, lowing alcohol consumption, lowing saturated fat in the diet, and regular cardiovascular exercise could all reduce the risks of incidence [18]. In addition, maintaining a healthy weight, especially after the menopause, reduces breast cancer risk.

#### 1.1.3 Prognosis

Prognosis provides an estimate of the likely course and outcome of a disease. Breast cancer prognosis is generally affected by several factors, including how advanced the tumour is (i.e. stages), the pathological characteristics (size, grade and stage) and the patient's general health [19]. Based on the location of the tumour, breast cancers can be classified into early, local advanced or secondary tissue [20]. Early breast cancer indicates tumours that are still confined to the breast and have not spread to other parts of the body. Local advanced breast cancer referrers to tumours that have spread to the surrounding area of the breast tissues, such as the lymph nodes in the armpit, skin or chest muscle. Local advanced breast cancers usually are larger compared to early breast cancers. Secondary breast cancer, also known as metastatic and describes tumours that have spread to the other part of the body through the lymphatic system or blood vessels. Common breast cancer metastasis sites include the bones, the lungs, the brain and the liver [21].

In the UK, the TNM (tumour, node and metastasis) staging system is the most commonly used by breast pathologists during biopsy and histological examinations and screening diagnosis [22]. The TNM system measures a patient's tumour size (T), lymph node stages (N) and metastasis status (M). These three parameters can be used to provide a prognosis estimation (i.e. 5-year survival rate) using the nottingham prognostic index (NPI), first devised by Galea et al. in 1992 [23]. The original NPI system gave a score that estimated a patient's prognosis as good, moderate or poor. The modern NPI system used in the UK is refined and segregates patients into four groups (excellent, good, moderate and poor) [24]. Breast cancers are also described as different *grades* and *stages*. The grading system is derived from a composite score, taking into account of nuclear pleomorphism (the nucleus size and shape of cancer cells), mitotic rate (how fast the cancer cells are multiplying) and tubule formation (ratio of cancer cells formed into tubules). The grade describes the degree of differentiation, with a lower grade indicating a welldifferentiated and slower-growing tumour and a higher grade indicating a poorly differentiated and faster-growing tumour. In general, lower-grade tumours are predicted to have a better prognosis than higher-graded tumours. Tumours are also defined by stages. Different from the grade, the stage describes the size and metastasis status of the tumour [25]. The number staging system in breast cancer divides breast cancers into four *stages*, with each stage having a series of subcategories. Stage 1 breast cancers are small (<2cm) and only found within the breast. Stage 2 breast cancers are still relatively small (2-5cm) but may have spread to the nearby lymph nodes. Stage 3 breast cancer is larger (>5cm) and has spread to other parts of the chest, such as local lymph nodes, skin or chest wall. Finally, stage 4 tumours are characterised by their large size and metastatic spread to nodes and distant sites. High-staged breast cancers are predicted to have worse prognosis than lower staged breast cancers.

Apart from the tumours' morphological features, the prognosis is affected by the tumours' biology. Specifically, tumours can be categorised by their expression status of druggable proteins, including oestrogen receptor alpha (ER $\alpha$ ), progesterone receptor (PR) and human epidermal growth factor receptor 2 (Her2). About 75% of the breast cancer express oestrogen receptor (note: US spelling 'estrogen receptor') (ER) $\alpha$  and/or PR, and these tumours are often referred to as ER-positive tumours [26]. Patients with ER-positive tumour generally have a better prognosis as the tumours tend to be less aggressive by nature and can be treated with targeted hormone therapy such as tamoxifen. A large study with a cohort of 111,993 patients showed that 90-95% of ER-positive survived longer than five years after diagnosis compared to 78-80%for ER-negative patients [27]. Other smaller studies [28–31] have all demonstrated similar results with ER-positive patients having a better 5-year survival rate compared to ER-negative patients. Another targetable protein in breast cancer is Her2, and about 15-20% of breast cancer express this protein. They are generally more aggressive and have a worse prognosis than ER-positive tumours. However, Her2-positive tumour can be treated with monoclonal antibody therapy (e.g. trastuzumab) and if the tumour responds to the treatment, the patient's prognosis is expected to improve [32]. Tumours that do not express  $ER\alpha$ , PR and HER2 are known as triple negative breast cancer (TNBC), and this phenotype accounts for about 10-15%of breast cancer cases. TNBC are highly aggressive and invasive tumours that tend to be higher grade than other types of breast cancer. Shown by several large studies, patients with TNBC have worse clinical outcomes than patients with other breast cancer types [33–35]. The overall five-year survival for TNBC is 62.1% compared to 80.8% for non-TNBC [36]. Patients with TNBC have also been linked to a high recurrence rate, especially within the first five years after the initial diagnosis [37]. The lack of three targetable proteins limits the treatment options for TNBC, contributes to the poor prognosis. Non-selective treatments (such as chemotherapy) remain the conventional treatment routine for TNBC. Treatment options are described in detail in Section 1.1.5. Other factors such as age, ethnicity, socioeconomic status, and the patient's general health all have an impact on the prognosis and clinical outcome [38].

#### 1.1.4 Molecular classification of breast cancer

Apart from clinical information and histopathological analyses, molecular-based classifications are also becoming an important approach to stratify cancer patients. With the pharmaceutical industry moving into the era of precision medicine (instead of systemic treatment) where the development of drugs, especially anti-cancer agents, is founded on biological mechanisms, it is important that the cancer stratification schemes take into account of the tumour's biological properties. The development of next-generation sequencing (NGS) techniques has allowed researchers to study the biology of breast cancer tumours from broader and deeper perspectives. The three largest molecular-based projects that have been performed so far are The Cancer Genome Atlas project (https://cancergenome.nih.gov), the International Cancer Genome Consortium (https://icgc.org), and the Molecular Taxonomy of Breast Cancer International Consortium (http://www.cbioportal.org/study?id=brca\_metabric). The extensive molecular profiling data accumulated by these projects help to improve the existing breast cancer classification and facilitate the discovery of novel drug targets.

The breast cancer molecular subtypes emerged as a result of the work done by Perou *et al.* [39] and Sorlie *et al.* [40], who analysed gene expression data taken from breast cancer samples and identified approximately 550 genes that reflect the phenotype of individual tumours. Based on the expression pattern of these genes, two main clusters of patients were identified (mostly reflecting the ER status) and were subsequently classified into five molecular subtypes: Luminal A, Luminal B, Her2, Basal-like and normal-like. The five molecular subtypes have also been found in samples from other independent cohorts. For example, Hu *et al.* validated the five breast cancer subtypes in a cohort of 105 samples, and their results suggested that these five subtypes could be distinguished by a unique a signature containing 306 genes [41]. Parker *et al.* reproduced the five molecular subtypes using a reduced 50-gene classifier (PAM50) [42]. It has been demonstrated further that the PAM50 subtyping scheme provides prognostic and predictive values [43, 44] and can provide a valuable contribution to the clinical setting [45, 46].

Sorlie's hierarchical subtyping scheme first stratified tumour samples into two large clusters. In most cohorts, the largest cluster is dominated by ER-positive tumours and tumours within this cluster generally express genes that are expressed by luminal breast epithelial cells (hence Luminal subtype). The tumours in this cluster can be classified further into Luminal A and Luminal B, with Luminal B tumours having a higher expression of genes involved in mitosis and proliferation. The other smaller main cluster is dominated by ER-negative tumours. Similar to the ER-positive cluster, ER-negative cluster can also be stratified into two subtypes: Her2-enriched and Basal-like. The Her2-enriched subtype is defined by overexpression of erbB2/HER2-related genes, whereas Basal-like breast cancer usually has a similar expression pattern to myoepithelial/basal epithelial cells (i.e. expressing cytokeratin 5/6). Basal-like subtype usually lacks  $ER\alpha$ , PR and Her2, demonstrating the triple negative phenotype. The fifth molecular subtype, normal-like breast cancer, has a similar expression pattern to normal breast tissue samples. However, there are doubts about whether normal-like breast cancer represents cancerous tissues. Some researchers believe that normal-like breast cancer is a technical artefact from high contamination with normal tissue during the microarrays [47]. In addition to the original five subtypes, a claudin-low subgroup was later identified [48]. Similar to Basallike subtypes, claudin-low tumours also lack expression of ER $\alpha$ , PR and Her2. However, their inconsistent expression of basal keratins and low expression of the proliferation marker ki67 makes them distinct from Basal-like breast cancer.

Table 1.1: Summary of the molecular subtypes derived from the work by Perou *et al.* and Sorlie *et al.* [39, 40]. IHC: immunohistochemistry. OS: overall survival. N/A: not available. Context adopted from Dai [49] and Holliday [50].

Subtypes	IHC status	Characteristics	OS rate $[51]$
Luminal A	ER+, PR+/-, Her2-	Ki67 low,	95.9%
		less aggressive	
Luminal B	ER+, PR+/-, Her2-	Ki67 high,	93.5%
		less aggressive	
Her2-enriched	ER-, PR-, Her2+	Ki67 high,	90.5%
		aggressive	
Basal-like	ER-, PR-, Her2-	Ki67 high,	82.6%
		Some EGFR+,	
		cytokeratin $5/6+$ ,	
		aggressive	
Claudin-Low	ER-, PR-, Her2-	Ki67 low,	89.7%
		Claudin- $3,4\&7$ low,	
		invasive (E-cadherin low),	
		aggressive	
Normal-like	ER+, PR+, Her2-	Ki67 low	N/A

While the five molecular subtypes discovered by Perou *et al.* and Sorlie *et al.* have set a standard for molecular subtyping, other classifications do exist. For example, using a signature containing 706 cDNA probe elements Sotiriou *et al.* identified six breast cancer subgroups, which include three Luminal, one HER2-enriched and two Basal-like subtypes [52]. In addition, Lehmann *et al.* analysed 587 tumours samples that displayed triple negative characteristics and subdivided them into six stable groups: two Basal-like groups (BL1 and BL2), one immunomodulatory (IM), one mesenchymal (M), one mesenchymal stem-like (MSL) and one Luminal androgen receptor (LAR) subtype [53]. Another study that analysed 995 breast tumour samples suggested that there are at least ten clinically distinct subgroups of breast cancer [54]. These clusters were revealed based on an integrated clustering method that considered multiple layers of data, including copy number variations, single nucleotide polymorphisms and gene expression patterns. The authors reported that in these ten clusters of breast cancer, the majority of Basal-like tumours formed a stable subgroup that displays high-genomic instability (high frequency of mutations).

#### 1.1.5 Basal-like breast cancer

The focus of the research presented in this thesis is on basal-like breast cancer (BLBC). As mentioned, BLBC often lacks expression for ER $\alpha$ , PR and Her2 and its triple negative characteristic has led to the common synonymic mistake by which triple negative breast cancer (TNBC) is considered to be as same as BLBC. It should be noted that, although TNBCs and BLBCs share many similarities, they are not identical. About 20% of TNBCs are not Basal-like by molecular profiling, and about 15-45% of Basal-like do not have TNBC characteristics (Figure 1.1). The Basal-like breast cancer is defined by a distinct gene expression pattern characterised by high expression of Basal markers (cytokeratin 5,6 and 17). Whereas triple negative breast cancer refers to the heterogeneous group of tumours that have negative immunohistochemical staining for ER, PR and Her2. As more is discovered about the biology of TNBC, it is clear that TNBC is not a single homogeneous tumour type. Indeed, at least seven subgroups of TNBCs have been identified, and among them, Basal-like has one of the worst prognoses [55]. A recent study by Dogra *et al.* that analysed medical records from 200 TNBC patients showed that Basal-like TNBC has the shortest disease-free survival and overall survival compared to non-Basal-like TNBCs [56].



Figure 1.1: Venn diagram representing the relationship between triple negative breast cancer and Basal-like breast cancer. Basal-like breast cancer is defined by positive staining of cytokeratin 5,6 and 17, and 15-45% of Basal-like breast cancer do not have triple negative characteristics. Triple negative breast cancer refers to tumours that lack expression of ER, PR and Her2, and 20-30% of TNBC are not Basal-like [57]. Figure reproduced from [58].

While the prevalence of BLBC varies from study to study, the poor prognosis of patients with Basal-like tumours has been repeatedly seen in most cohorts [40, 52, 59–64]. In particular, Basal-like breast cancer is associated with shorter 5-year survival, high risk of early relapse (2-5 years after treatment) and the development of distant metastasis [65]. Different from other breast cancer subtypes, Basal-like is mostly found in young and premenopausal women, especially among African American women [66]. Basal-like has also been associated to a specific pattern of distant metastasis with a tendency to distribute to the brain and lungs, and less likely to distribute to the to bone and the liver, suggesting that Basal-like tumours have a distinct mechanism of metastatic spread [67, 68].

For more than two decades, the status of ER, PR and Her2 have been used in the clinical practice to facilitate prognosis prediction and treatment design. Since the development of molecular classification, these three markers have been used to approximate the molecular subtype of tumours. However, this over-simplified approach classifies all TNBC into the Basal-like category. To clarify the disparity between TNBC and Basal-like breast cancers, a number of biomarker candidates have been proposed in an attempt to better identify Basal-like in the clinic. Apart from creatine kinase 5 (CK5) and creatine kinase 6 (CK6) that are expressed in approximately 73% of patient biopsies, other identified biomarker candidates include insulin-like growth factor II mRNA binding protein 3 (IMP3)(79%), vascular endothelial growth factor (VEGF)(78%), p53 (62%), epidermal growth factor receptor (EGFR)(50%), creatine kinase 14 (CK14) (44%), creatine kinase 17 (CK17) (33%), stem cell growth factor receptor Kit (KIT) (31%) [69–71]. Among them, only basal CKs (CK5, CK6, CK14 and CK17) have been shown independently to be associated with poor clinical outcome, and the other biomarkers have not added value to identifying Basal-like cases compared using the creatine kinase (CK)s alone [57].

While previous studies have shown that Basal-like has distinctive genetic, immunophenotypic, and clinical features, to date, there is not one internationally accepted definition for this breast cancer. In this work, the Basal-like breast cancer refers to the Basal-like subtype defined by Parker's PAM50 classification scheme [42].

### 1.1.6 Treatments

With the latest advances in molecular biology and high-throughput technology, several molecular therapeutic targets have been identified. Their discovery could hail a new generation of anti-cancer drugs that are tailored according to the specific molecular and pathophysiological features or phenotypes of cancers. The application of targeted therapy in the clinical settings has significantly improved the survival rate for breast cancers, with more than 75% of breast cancer patients are predicted to survive ten years or longer in the UK in 2019 compared to 40% in the 1970s. Apart from targeted therapy, surgery, radiotherapy followed by the more traditional (non-selective) chemotherapy are still used routinely, in combination in the clinic.

Targeted treatments are defined by tumour gene expression profiles. To assess whether a patient is going to respond to targeted treatments, breast cancers are classified by therapeutic clinical subtypes: ER-positive, HER2-positive and triple negative breast cancer (TNBC). ER-positive refers to breast cancers that express oestrogen receptor alpha (ER $\alpha$ ). Oestrogen receptors are activated by oestrogen, a key driver for breast cancer progression. Oestrogen receptors can be targeted by selective oestrogen receptor modulators (SERM), also called hormone therapy that inhibits the ER-related signalling pathway. Inhibition of oestrogen signalling pathway has been shown to suppress tumour growth and reduce the relapse rate in ER-positive breast cancers [72]. Tamoxifen was the first SERM to be approved to treat ER-positive breast cancers. Tamoxifen acts as an antagonist to ER $\alpha$ , preventing oestrogen from activating its receptor and inhibiting proliferation. Treating patients with tamoxifen after surgery has been demonstrated to reduce the recurrence rate by approximately 40–50% in women with early breast cancer [73]. Tamoxifen has also been shown to have preventive effects in women at higher risk for developing breast cancer, reducing overall breast cancer incidence up to 49% [74]. Other SERMs such as raloxifen [75–77] and lasofoxifene [78] have also entered clinical trials.

Aromatase inhibitors (AIs) represent another class of targeted therapy for ER-positive breast cancers. AIs were developed to reduce the circulating oestrogen levels in the body by blocking the aromatase enzyme from converting androgen to oestrogen. Examples of the AIs include letrozole, anastrozole, and exemestane.

The Her2 receptor is the most commonly overexpressed receptor in breast cancer and is also considered an efficient drug target. Tumour with Her2 overexpression, Her2-positive tumours, are treated with targeted treatment in the form of a Her2 specific monoclonal antibody therapy. The recombinant antibody trastuzumab (Herceptin) was the first approved drug to treat Her2positive breast cancers. It has been demonstrated that treating Her2-positive patients with trastuzumab has significantly improved disease-free and overall survival in the early stage cancer [79]. While trastuzumab is widely used to treat Her2-positive tumour, its exact molecular mechanism of action remains unclear. The most popular explanation is that it targets the extracellular domain of Her2 and prevent Her2 receptors from dimerising which is crucial for activating the downstream growth factor signalling cascades including the phosphatidylinositol 3-kinase (PI3K), protein kinase B (AKT), mechanistic target of rapamycin (mTOR) pathway [80–82]. By downregulating the PI3K, AKT, mTOR pathway, trastuzumab induces cell cycle arrest and consequently inhibit proliferation. The discovery of trastuzumab was followed by other agents like pertuzumab and lapatinib, both approved by the food and drug administration (FDA) in the USA [83–85].

Triple negative breast cancer is the most heterogeneous therapeutic clinical subtypes. It has a higher risk of recurrence and a shorter overall survival rate compared to ER-positive and Her2-positive. To date, anthracyclines, taxanes and anti-metabolites are the FDA-approved chemotherapy regimens recommended for TNBC, in both adjuvant and neoadjuvant settings. Unfortunately, due to the lack of efficiency drug targets, there is currently no available targeted therapy for TNBC, and the standard chemotherapy is still the conventional route to treat TNBC patients. Bevacizumab, a monoclonal antibody against VEGF, was initially approved by FDA for TNBC but then withdrawn because of the disappointing effect on the overall survival when used in combination with chemotherapy [86, 87]. Other agents that have shown disappointing clinical results in TNBC includes vascular endothelial growth factor receptor (VEGFR) antibody ramucirumab [88], tyrosine kinase inhibitor sunitinib [89, 90] and sorafenib [91], EGFR antibody cetuximab [92, 93], small-molecule tyrosine kinase inhibitors gefitinib [94] and erlotinib [95]. On the positive side, the Poly ADP-ribose polymerase (PARP) inhibitor Olaparib has shown a favourable objective response rate in TNBC tumours with both BRCA1 and BRCA2 mutations [96].

As mentioned before, Basal-like breast cancer is aggressive, and as a subset of TNBC, there is currently a vast scope for new and improved treatment options, particularly highly effective targeted therapies. With most of the previous drug agents demonstrating disappointing results, there is indeed a priority to identify new molecular targets for Basal-like breast cancer and other subsets of TNBCs. This research will reveal new insights into the transcriptional characteristics of Basal-like breast cancer to suggest new drug target candidates for Basal-like breast cancer.
# **1.2** Nuclear receptors

nuclear receptors (NRs) have been an emphasis of many cancer drug development research strategies, due to their roles as critical regulators of diseases. In breast cancer, oestrogen receptor and progesterone receptor, both are members of the nuclear receptor superfamily, remain to be clinically important biomarkers for predicting prognosis and determining therapeutic strategies. More recently, roles of other NRs have also been explored in breast cancers, and there is growing evidence suggesting the involvement of multiple nuclear receptors other than ER and PR in breast cancer. With the development of genomic technologies, researchers are now able to profile the genomic locations and activation of NR in a genome-wide manner. Such studies have improved our understanding of NR activities in oncogenesis. In this work, the nuclear receptor genomic data were used to reveal the underlying NR expression patterns in Basal-like breast cancer and provide insight into the complex interplay of NR transcriptional networks.

# 1.2.1 Nuclear receptor superfamily

The human nuclear receptor (NR) superfamily consists of 48 members. They are transcriptional factors and responsible for regulating many physiological processes, including metabolism, immunity, developmental patterning and cell proliferation. This superfamily was discovered due to their highly conserved structural organisation [97], with most containing four major domains: firstly, the N-terminal regulatory domain with activation function 1, secondly, the highly conserved DNA-binding domain consisting of two zinc fingers that bind to the hormone response elements in DNA, thirdly, the hinge region linking the DNA-binding domain to the final ligandbinding domain, that can have an activation function 2 region that mediates coregulator interaction. Unlike other transcriptional factors, most nuclear receptors are activated by ligands. The common ligands for NRs include steroid hormones, thyroid hormones, lipophilic vitamins and cholesterol metabolites including retinoic acid and oxysterols [98]. However, about half of NRs have no known endogenous ligands and are referred to as orphan receptors or adapted orphans. According to their sequence homology, nuclear receptors can be divided into seven families (Table 1.2) [99]. Within each family, they can be further classified into groups depending on their associated ligands and the molecular pathways involved (Table 1.2) [100, 101]. More recent evolutionary groups based on their sequence alignment and phylogenetic tree construction have also been proposed [102].

rom Weikum [103] and [99]					
Family		Subgroup	Ligand	Members	Associated Biological Pathway
	Α	Thyroid hormone recep-	Thyroid hormone	NR1A1,	This family contains many orphan
		tor		NR1A2	receptors, and some have been shown
	В	Retinoic acid receptor	Retinoic acid	NR1B1,	to bind fatty acids. However, it is
L. I hyroid Hormone				NR1B2,	unclear if these binding events lead to
Receptor-like				NR1B3	ligand-driven regulation. Retinoid X
	C	Peroxisome	Fatty acids	NR1C1,	receptors form heterodimeric
		proliferator-activated		NR1C2,	complexes with other NRs and are
		receptor		NR1C3	the only receptors in this family with
	D	${ m Rev-ErbA}$	Heme	NR1D1,	a known activating ligand.
				NR1D2	
	Ц	RAR-related orphan re-	Sterols	NR1F1,	
		ceptor		NR1F2,	
				NR1F3	
	Η	Liver X receptor-like re-	Oxysterols, Bile acids	NR1H2,	
		ceptor		NR1H3,	
				NR1H4	

Table 1.2: Nuclear receptor families. SF1: Steroidogenic factor 1. LRH1: Liver receptor homolog-1. DAX: Dosage-sensitive sex reversal, advanal hypothesia critical region on chromosome X gene 1. SHP. Small beterodimer partner LRD. ligand-binding domain. Context advarted adrens

	Ι	Vitamin D receptor-like	Vitamin D, xenobiotics,	NR111,	
		receptor	androstane	NR112,	
				NR113	
	Α	Hepatocyte nuclear	Fatty acids	NR2A1,	Receptors in this family are regulated
		factor-4 receptor		NR2A2	by a range of lipophilic signalling
II. Retinoid X Receptor-like	В	Retinoid X receptor	Retinoids	NR2B1,	molecules, including thyroid
				NR2B2,	hormone, fatty acids, bile acids, and
				NR2B3	sterols.
	U	Testicular receptor	(Orphan)	NR2C1,	
				NR2C1	
	딘	TLX/PNR	(Orphan)	NR2E1,	
				NR2E3	
	Гц	Chicken ovalbumin up-	(Orphan)	NR2F1,	
		stream promoter recep-		NR2F2,	
		tor		NR2F6	
	Α	Oestrogen receptor	Oestrogens	NR3A1,	This family comprises steroid
III. Oestrogen Receptor-like				NR3A2	receptors. They bind to
	В	Oestrogen related re-	(Orphan)	NR3B1,	cholesterol-derived hormones and
		ceptor		NR3B2,	regulate metabolic, reproductive, and
				NR3B3	developmental processes.

				Members of this family are responsi-	ble for neuron development and main-	tenance.	Members of this family are responsi-	ble for neuron development and main-	tenance.	NR6A1 remains in its own family	due to its unique characteristics in	the LBD (ligand-binding domain).	NR6A1 has a critical role in neuroge-	nesis and germ cell development.
NR3C1,	NR3C2,	NR3C3,	NR3C4	NR4A1,	NR4A2,	NR4A3	NR4A1,	NR4A2,	NR4A3	NR6A1				
Cholesterol-derived	hormones			Unsaturated fatty acids	(NR4A2) or $(Orphan)$		Unsaturated fatty acids	(NR4A2) or $(Orphan)$		(Orphan)				
3-Ketosteroid receptors				Nerve Growth Factor	IB-like		Nerve Growth Factor	IB-like		Germ cell nuclear factor				
G				IV. Nerve Growth Factor IB- A	like		V. Steroidogenic Factor-like A			VI. Germ Cell Nuclear A	Factor-like			

Miscellaneous	В	DAX/SHP	(Orphan)	NR0B1,	This family includes atypical NRs.
				NR0B2	NR0B1 and NR0B2 are different from
					other NRs and contain only an ligand
					binding domain, which allow them to
					interact with other NRs to regulate
					transcription.

Nuclear receptors are signal proteins responsible for transducing cellular signals. Upon ligand activation, NRs typically undergo conformational changes that allows them to bind to their corresponding DNA targets and regulate gene transcription [104]. NRs can bind to the target DNA as monomers, homodimers or heterodimers with another family member. For example, oestrogen receptors and other steroid hormone receptors generally bind as homodimers, whereas non-steroid hormone receptors including vitamin D3 receptor (VDR), thyroid hormone receptor (THR), retinoid receptor (RAR) and retinoid X receptor (RXR) form both homodimers and heterodimers. RXR are usually the heterodimer partners for VDR, THR, RAR and orphan receptors. NR dimerisation is an essential biological process, providing specificity and combinatoric diversity.

Ligand binding is also an important process that controls NR functionality and activities. It has been shown that, in the absence of ligand, NRs tend to be conformationally unstable, and ligand binding increases the stability of the ligand-binding domain [105, 106]. The stabilisation provided by the ligand binding facilitates interactions with coregulators [107]. NR coregulators often function as a member of a large multi-protein complex. By interacting with NRs and other transcriptional factors, coregulators facilitate or inhibit transcription of the corresponding target gene into mRNA. Apart from DNA, ligands and coregulators, NRs have also been found to complex to other transcription factors on the chromatin [108].

## 1.2.2 Coregulators

As mentioned before, NRs are ligand-induced transcription factors that regulate gene expression, controlling a wide range of essential cellular process. Similar to other transcription factors, ligand-bound NRs recruit RNA polymerase II or other proteins to the chromatin environment. The proteins that are involved in a complex that can be recruited by nuclear receptors to regulate the transcription of target genes are known as coregulators. Most nuclear receptor coregulators function in large multi-protein complexes and assist in regulating NR-mediated transcription by modifying chromatin accessibility, stabilising the NR–DNA interactions or facilitating indirect NR-DNA interaction [108, 109].

The first NR regulators were discovered by identifying proteins that directly interacted with the NRs [110–112]. These protein partners are called the core coregulators or the primary coregulators. Two classes of core coregulators have been identified: the coactivators and the corepressors. Coactivators are proteins that when recruited to the chromatin will activate the transcription of the target gene. In contrast, corepressors are those that repress the transcription of the target gene. Coregulators are typically, but not always, recruited in a ligand-dependent fashion. In most cases, coactivator proteins are recruited to ligand-bound NRs; whereas corepressor complexes are recruited to unbound NRs. Coregulators do not have to make direct contact with NRs to affect NR-mediated transcription. The protein complexes that influence the rate of NR-mediated transcription, but do not directly interact with the NRs, are known as secondary coregulators. Importantly, coregulators can be recruited to more than one transcription factor; and similarly, transcription factor can often recruit more than one coregulator complex. Moreover, the role of a coregulator protein complex is not fixed, and depending on the cellular and signalling context, coactivators and corepressors can sometimes switch roles and have an opposite effect on the transcription.

The associations between NRs and their coregulators are tightly regulated, and small changes (e.g. mutations, miss-expression, abnormal post-translational modifications) could have a dramatic impact on the functionality of NRs and coregulators, potentially give rise to human diseases [113]. In a cancer transcriptome meta-analysis performed by Rhodes *et al.*, it showed that NR coregulators are often over- or under-expressed in many cancers, suggesting misexpression of coregulator is likely to contribute to cancer progression [114]. In breast cancer, alternation in coregulator expression has been implicated in hormone therapy resistance [115], suggesting expression level of coregulators could be valuable for determining the tissue-specific response to hormone therapy such as tamoxifen [116, 117]. Moreover, it has been shown that coregulator show differential expression in breast cancer compared to normal breast tissues [118].

Taken together, coregulators are a critical aspect of NR-mediated transcriptional regulation in both physiological and pathological conditions. Furthermore, the expression levels of coregulators can determine tumour responsiveness to hormone therapy. Targeting coregulators, either independently or in combination with NR inhibitors, represent a potential treatment strategy for treating breast cancer. In this research, bioinformatics approaches were used to reveal the expression patterns of NRs and their core coregulators in breast cancers, specifically, to investigate whether the expression patterns are different in different breast cancer subtypes, with an inevitable emphasis on Basal-like breast cancer.

## 1.2.3 Involvement in breast cancer

In the context of breast cancer, the roles of ER and progesterone receptor (PR) have been intensively studied. ER and PR have been shown to be involved in the progression and development of breast cancer, and their expression status remains a robust prognosis indicator. In addition, androgen receptor (AR) has recently emerged as a clinical outcome biomarker and a potential drug target [119, 120]. Other NRs such as vitamin  $D_3$  receptor (VDR) [121, 122] and glucocorticoid receptor (GR) [123, 124] have also been shown to affect cell proliferation and apoptosis in breast cancer. Furthermore, members of retinoid receptor (RAR) subfamily have also been investigated. The effects of retinoids (RAR ligand) is highly dependent on the presence of ERs [125], suggesting there is a crosstalk between RARs and ER signalling pathways. Retinoic acid receptor beta has also been shown to have an anti-migratory effect in breast cancer cells [126, 127]. Figure 1.2 summarised the NR-mediated signalling pathways that have been implicated in breast cancer.



Figure 1.2: NR-mediated signalling pathways that have been implicated in breast cancer Nuclear receptor and their corresponding ligands are shown in colour. Non-NR signalling proteins are shown in grey. AR: androgen receptor; PI3K: phosphoinositide 3-kinase; SRC: proto-oncogene tyrosine-protein kinase Src; MAPK: mitogen-activated protein kinase; AKT: protein kinase B; mTOR: mechanistic target of rapamycin; ER: estrogen receptor; PR: progesterone receptor; RAS: Ras GTPase; Raf: RAF kinases; MEK1/2: mitogen-activated protein kinase kinase; GR: glucocorticoid receptor; Wnt: wingless and Int-1; GSK-3 $\beta$ : glycogen synthase kinase  $3\beta$ ;  $\beta$ -catenin: catenin beta-1.

There is increasing evidence to suggest that the involvement of NRs in breast cancer extends beyond regulating proliferation and apoptosis. NRs are involved in other aspects of breast cancer tumorigenesis, such as regulating the circadian clock, aspects of energy metabolism and metastasis. The circadian clock is a biological process that that regulates the sleep-wake cycle and is coordinated with solar time (typically repeats every 24 hours). Circadian clock disruption (e.g. night shift workers) has been linked to an increased risk of breast cancer [128]. While circadian clock initially refers to a relatively simple feedback loop mechanism involving transcription factors CLOCK, BMAL and their coregulators, it is now emerging that the circadian clock contains additional loops that involve a number of NRs. To date, RARrelated orphan receptors [129], NR1D1 [130], glucocorticoid receptors [131] and oestrogen related receptor alpha (ESRRA) [132] have all been demonstrated to have a regulatory role in controlling the circadian rhythm.

In addition, several NRs, such as REV-ERBs [133, 134], oestrogen-related receptors [135, 136] and peroxisome proliferator-activated receptors (PPARs) [137], have been shown to be involved in energy metabolic pathways and their abnormal expression in breast cancer is thought to give rise to metabolic reprogramming of the tumours cells [138] (REV-ERBs are summarised by De *et al.* [139]; oestrogen receptor-related receptors (ERRs) are summarised by Deblois *et al.* and Cai *et al.* [136, 140]; PPARs are summarised by Sakharkar *et al.* and Avena *et al.* [141, 142]).

Finally, NRs have also been shown to be involved in cellular pathways that influence breast cancer cell migration and metastasis. For example, oestrogen receptor signalling pathways have been shown to influence cell migration and invasion, with oestrogen receptor 1 (ER1) inducing cell migration and oestrogen receptor 2 (ER2) having the opposite effects [143]. Moreover, activating PRs using progestins, synthetic forms of progesterone, have also been shown to have an anti-migration effect in barest cancer [144, 145]. These effects however are ligand and PR isoform dependent. The involvement of other NRs, such as retinoic acid receptors [127] and oestrogen-related receptors [146], in influencing the breast cancer invasive and metastatic behaviour have also received research attention and are being investigated.

In summary, studies have highlighted the involvement of NRs in regulating multiple pathways, demonstrating the complexity of NR signalling. Dysregulation of NR signalling can have a significant impact on breast cancer progression and tumorigenesis. Given their significant involvement in pathology and their ligand-activated nature, NRs have emerged as a valuable class of therapeutic targets [103]. Understanding the functions and transcriptional dynamics of NR and their coregulators will open up more opportunities for targeting oncogenic pathways modulated by NRs.

# 1.3 Computer-aided drug discovery

Computational approaches have been applied in different steps of the drug discovery process. The appreciation and popularity of computational methods continue to grow due to the advances in computational architectures and technologies. Drug discovery is a costly and time-consuming process, which takes, on average, 12-15 years and typically \$1.3 billion to bring a new drug into the market for treating patients [147]. The cost of failures hugely contributes to these figures. It is estimated that 80 to 90% of the research projects fail before they enter clinical trials [148]. There is an urgent need to reduce the cost and timeframe for the drug discovery process. The application of computational strategies contributes to bringing down the cost and time needed as well as reducing the failure rate [149], and due to this, computer-aided drug design has become an essential part of modern drug discovery.

#### 1.3.1 Drug discovery process

Drug discovery is described as the process of developing therapeutic agents for the treatment and management of diseases. Historically, discoveries of drugs were more serendipitous, with most compound, such as penicillin, identified from natural extracts or traditional remedies [150]. The modern drug discovery represents a systematic process, involving multiple steps and a considerable amount of research effort from multiple scientific disciplines (Figure 1.3).



Figure 1.3: Stages of the drug discovery process.

Modern drug discovery process begins with understanding the molecular and mechanistic pro-

cesses in diseases. This understanding allows scientists to identify potential targets for intervention. Typically, drug targets are large biological molecules (i.e. RNAs, proteins) that are involved in the pathology of interest. With the advances in sequencing technologies, omics approaches are used to identify the underlying expression patterns implicated in diseases from large datasets [151, 152]. Target identification is followed by target validation, which is the process of characterising the functionality of the presumed target in the disease phenotype. Functional screening and pathway analysis can be examined experimentally using knockdown, mutants or computationally by data mining. Other validation processes include evaluating the effect of mutations, knockdown or overexpression on the target of interest to examine the known signalling pathways downstream of the targets of interest. Computational validation is often followed by experimental experiments to ensure its reproducibility.

Once a target is identified and validated, the next step of the drug discovery process is to identify a chemical series based on common scaffolds or based on low molecular weight fragments that give candidate lead compounds. Lead compound are synthetically accessible and feasible chemical series that have pharmacological and biological activity likely to be therapeutically useful. During this process, a series of synthesis is generated based on the lead molecule, providing an indication of the biochemical and structural characterisation requirement for a molecule to bind to the desired target. Lead optimisation aims to maintain the desirable features in the lead compounds while improving the deficiencies. The improvement is often achieved by altering the chemical structure of the lead compound.

To increase the success rate of the drug development process, druggability assessment is often performed. This assessment examines the likelihood a compound binds to the desired target. Other examinations include pharmacokinetic profiling, metabolism and toxicity are often conducted to evaluate the degradation and elimination of a compound [153, 154]. The common properties evaluated during the optimisation process includes efficacy, specificity, selectivity, toxicity, stability and bioavailability.

When a chemical compound demonstrates promising therapeutic activities, the next step in the drug discovery process is to characterise the physicochemical properties of the drug candidates. In this stage, the possible mechanisms for the best delivery system are investigated to ensure drug stability and bioavailability under physiological conditions.

Pre-clinical research evaluates a drug's safety and efficiency both in vivo and in vitro, for

selecting the most appropriate drug candidates for human clinical trials. Pharmacology and toxicology are two critical aspects of pre-clinical evaluation [155]. Pharmacology studies explore the pharmacokinetic and pharmacodynamic behaviour of the drugs, crucial for understanding the absorption rate for different routes of administration, distribution in various part of the body, rate of metabolism and finally, excretion. All of these properties are useful for determining the half-life for a drug. Half-life is defined as the time that takes a drug to reduce its concentration by 50% in the body. Toxicological studies profile the toxicity of a drug both *in vitro* and *in vivo*. *In vitro* studies evaluate how a drug influence cell behaviour such as proliferation. Whereas *in vivo* studies evaluate the toxicological effects on laboratory animals. The pharmacological and toxicological profiles and insight gained from this stage are used to form the foundation of clinical trials.

Compounds showed favourable pharmacological, and toxicological properties will enter clinical trials. Clinical trials are performed in volunteers, to answer specific research questions such as the safety or efficacy of a drug compound. There are typically 5 phases of clinical trials from phase 0 to phase IV.

This brief introduction into the drug discovery process has illustrated that discovering and developing a new drug compound is a complex, time-consuming and expensive task. Advances in data-driven and computational approaches have improved the efficiency, productivity and research capacity significantly at many stages of the drug discovery pipeline, reducing the time and cost of the process [156]. Specifically, in this following work, a computational pipeline was developed to identify potential drug targets by revealing the unique transcriptional pattern implicated in Basal-like breast cancer from two large publicly available datasets.

## 1.3.2 Bioinformatics in drug target identification

As mentioned earlier, drug target identification is a crucial early step in the drug discovery process that aims to identify biomolecules that could be modulated to provide pharmaceutical benefits [157]. Typically, putative drug targets are identified based on the existing scientific literature and/or biomedical databases, relying on collecting and integrating data from multiple sources [158]. Rapid development of the high throughput technologies has allowed quantification of many biomolecules (i.e. RNA, protein), expanding the volume and diversity of scientific data available for this challenging task. Bioinformatics is an interdisciplinary branch of computational biology that explores and analyses big data from genomic, transcriptomic and proteomic sources. Bioinformatics approaches are now used to understand the pattern from molecular mechanisms underlying complex disease states; and to inform the development of a drug target hypothesis, representing an important stage of drug target identification [159].

In bioinformatics studies, genes that are (i) differentially expressed in disease or; (ii) coexpressed with genes that presumed to be involved in disease-associating pathways or; (iii) known to be involved upstream of the disease-associating pathways are often further investigated to explore their potential as drug targets. In the past two decades, studies have curated and characterised the biomolecular targets that entered the market [160, 161]. It has become clear that the most successful drug targets share several common characteristics, including involvement in a disease-associated pathway, the ability to respond to common therapeutic molecules such as small molecules, antibodies, proteins or peptides and finally, and are functionally and structurally well characterised. In terms of their biological classes, a drug target can be DNA, RNA, proteins etc. Some of the most frequently targeted proteins include proteases, kinases, G protein-coupled receptors and nuclear receptors [162]. Drug targets can also be categorised according to the biological processes, such as enzymes, receptors, transport proteins [163].

The development of bioinformatics has allowed scientists to better understand the molecular mechanism underlying diseases. Bioinformatics studies are not only useful for revealing drug targets but also beneficial for understanding the molecular basis of diseases [164]. As well as focusing on a particular class of targets, another strategy focuses on a specific disease or disease phenotype. This strategy uses gene expression profiles generated from microarray experiments to identify disease-relevant genes and proteins that can be targeted. Network-based approaches are also popular in bioinformatics to examine the relationships or interactions between biomolecules. Gene networks and protein-protein interaction networks are commonly used. Gene networks allow an understanding of how alterations impact the regulation activities and pathways. Alternatively, instead of representing regulatory relationships protein-protein interaction networks represent biophysical contacts. Interactions between proteins are concerted and typically serve a specific function. Revealing protein topological properties are useful in identifying crucial proteins central to signalling or metabolic pathways. In this study, an integrated bioinformatics strategy combining all of these four themes was developed. Particularly, this research focussed on revealing unique nuclear receptor expression pattern in Basal-like breast cancer, in which both gene correlation networks and protein-protein interaction networks have been used to prioritise drug target candidates. Table 1.3 summarises the bioinformatics databases and platform used in this study.

Database	Website	Description
cBioPortal - TCGA - METABRIC	https://www.cbioportal.org/ ~ study/summary?id=brca_tcga_pub ~ summary?id=brca_metabric	cBioPortal is a database containing large scale genomics data for more than 42 types of cancer. Two datasets, TCGA and METABRIC, were consid- ered in this research.
STRING	https://string-db.org	STRING database contains known and predicted protein-protein interac- tions.
UniProtKB	https://www.uniprot.org/help/ uniprotkb	UniProtKB is a protein database, containing functional, biological and structural information and annota- tions.
IntAct	https://www.ebi.ac.uk/intact/	IntAct is a curated protein interaction database.
MINT	https://mint.bio.uniroma2.it	MINT is a functional protein interac- tion database
PDB	https://www.ebi.ac.uk/pdbe/	PDB is a database for crystallographic or NMR 3D structural data for pro- teins and other large biomolecules.
SMART	http://smart.embl-heidelberg.de	SMART is used to identify protein do- mains within proteins.
САТН	https://www.cathdb.info	CATH is a protein structure classifica- tion database that provides informa- tion about the evolutionary relation- ships of protein domains.
Gene3D	http://gene3d.biochem.ucl.ac. uk/Gene3D/	Gene3D is a database that assigns CATH structural domains within pro- tein sequences.
ELM resource	http://elm.eu.org	ELM is a database that predicts short linear functional motifs within protein sequences.
GDSC	https://www.cancerrxgene.org	GDSC is the largest open-source database that provides information on drug sensitivity in cancer cells and molecular markers of drug response.

Table 1.3: Databases and used in this project. All databases are open-source and freely available online.

With most of the data and software used in this thesis being free and publicly available, the strategy developed in this PhD research provides an efficient and economical approach for target identification. It should be pointed out that while bioinformatics strategies are useful for identifying putative drug targets, laboratory-based drug target validation is required to examine whether manipulating these putative targets is likely to give rise to the rapeutic benefits.

# 1.4 Chapter summary

The aim of this PhD research is to consider the transcriptional expression patterns underlying Basal-like breast cancer and identify putative drug targets for this particularly aggressive subtype. The current molecular characterisation of BLBC has not yet led to any successful development of targeted treatments, and therefore there is an imperative need for the identification of new molecular targets for this subtype. The strategy developed in this research integrated work flow of computational biology and biochemistry approaches and can be easily adapted to study other diseases or drug target classes.

The work described in Chapter 2 aimed to investigate whether Basal-like tumours have distinctive NR expression patterns from other breast cancer subtypes. A model-based clustering method was used to stratify breast cancer samples according to their nuclear receptor transcriptional variation. Clustering analysis was followed by network modelling, which was used to examine the associations (partial correlations) between NRs in Basal-like breast cancer. In this chapter, a number of network-based modules that are enriched in Basal specific networks were identified, representing important transcriptional components underlying Basal-like breast cancer.

Patient-expression clustering analysis was performed in two directions. While Chapter 2 presents one set of the clustering results where patients were stratified according to the distribution of gene expressions, Chapter 3 describes the results from the other dimension, in which nuclear receptor genes were clustered according to information content from the patient profiles. From the gene clustering result, a small number of genes were found to have exceptionally high expression in Basal-like subtype compared to other subtypes. To prioritise these genes, their topology properties were examined using a correlation network, from which lymphocyte-specific protein tyrosine kinase (LCK) and signal transducer and activator of transcription 1 (STAT1) were ranked the highest and investigated further.

LCK and STAT1 mRNA expressions were assumed to be the first proxy for protein expression, and Chapter 4 explores the protein interactions surrounding LCK and STAT1, represented in a protein-protein interaction network. Protein KIT and GRB2 associated binding Protein 1 (GAB1) were central to the network, and the twelve protein-protein interactions bridging between them were further investigated. A subsequent structural bioinformatics analysis was performed to reveal and characterise their putative binding sites.

Results from Chapter 4 suggested that KIT and GAB1 shared a number of binding partners, including PI3K regulatory subunit 1 (PIK3R1), PI3K regulatory subunit 2 (PIK3R2), PI3K regulatory subunit 3 (PIK3R3). To examine whether KIT-PI3K and GAB1-PI3K protein-protein interactions invoked the possible binding sites revealed in the previous chapter, molecular docking was performed and described in Chapter 5. The results from the docking analyses supported the findings described in the previous chapter. This analysis also predicted the preferred binding conformations for the complexed structures, which was summarised to provide structural insight into the molecular recognition of KIT-PI3K and GAB1-PI3K protein-protein interactions.

Chapter 6 describes a laboratory-based target validation using cellular assays. Immunofluorescence assays revealed the expression pattern of the endogenous KIT and PI3K regulatory subunits in four different breast cell lines. Results from the subsequent proximity ligation assays confirmed that the critical protein-protein interactions identified from the protein network do exist in Basal-like cell lines, suggesting that these protein interactions are candidate drug targets. Finally, the potential effects of manipulating these PPI targets in Basal-like breast cancer were discussed speculatively.

Chapter 7 discusses the key findings of this research and expands on the the general limitations and future perspective of the work presented in this thesis.

# Chapter 2

# **Breast Cancer Subtype Clustering**

# 2.1 Introduction

In this chapter, a model-based hierarchical clustering method was used to stratify breast cancer patients into groups according to their nuclear receptor (NR) transcriptional variation. A network modelling approach was then used to describe and examine the NR-associated transcriptional activities underlying Basal-like breast cancer. Network-based modules that are enriched in Basal specific networks were identified. These network-based modules provided a more mechanistic understanding of essential transcriptional components underlying Basal-like breast cancer.

# 2.1.1 Patient clustering in cancer studies

The development of NGS technologies has allowed researchers to gain a deeper understanding of the genome activities underlying diseases. In cancer research, the NGS data have been mainly used to (i) reveal molecular signatures associated with clinical outcome (e.g. prognostic marker), (ii) identify mutated or differentially expressed genes, and (iii) identify heterogeneity in cancers and inform the development of cancer subtype classification. Clustering is a standard method for NGS data interpretation. Clustering analysis aims to group objects (e.g. patient tissue) according to their levels of similarity (e.g. gene expression patterns). Clustering analysis was first employed to reveal the diversity in gene expression among the blood cancer tumour samples and successfully identified clinically significant cancer subtypes [165, 166]. Ever since, clustering approaches have been applied to define subtypes in several cancers, including breast cancer [40, 167, 168]. A state-of-the-art example in breast cancer classification is PAM50 subtyping scheme proposed by Parker *et al.*. As mentioned in Chpater 1, this scheme classifies breast cancer patient into one of five subtypes based on the gene expression pattern of 50 genes [42]. While the PAM50 classifier provides a stepping-stone in developing a classification for breast cancer, research has reported ambiguities in the PAM50 subtyping and suggested that the molecular subtypes can represent indistinct or incoherent sample groups [47, 169]. This highlights the ongoing challenges of reproducibility and ambiguity in cancer classification. To address these issues, and improve the robustness of patient clusters, several clustering algorithms that take intrinsic characteristics (e.g. the high dimensionality) of gene expression data into consideration have been proposed [170–172]. In this chapter, a model-based clustering method was used to reveal a natural breast cancer stratification. This is distinct from other conventional hierarchical methods.

#### 2.1.2 Model-based clustering vs classical clustering

While clustering analysis can be useful in stratifying patients based on their genetic variation and inform targeted therapies, the most commonly used clustering algorithms typically make a bias assumption when determining the number of clusters. For example, the most commonly used clustering approach using hierarchical clustering algorithms identifies the number of clusters by inspecting the dendrogram produced, a tree-like diagram showing the bottom-up hierarchy of objects formed by step-wise fusions based on their similarity [173-175]. The vertical axis is a scaled fusion distance between objects. K-means clustering is another commonly used method, that requires a pre-defined number of clusters prior clustering. This semi-supervised approach is based on a pre-defined number of expected clusters. Another subjective decision the user often has to make when performing classical cluster analysis is the choice of a metric or norm, especially in preparing the pairwise metrix This is often overlooked and impacts the resulting analysis. A metric is a measurement that indicates the size of dissimilarity between any two objects (from 0 to infinity). This is typically a calculated distance in n dimensional space representing orthogonal variables. However, the similarity between any two objects is considered as a norm space, constrained by a coefficient from 0 to 1.0, and is commonly exemplified by the use of percentage or ratio. In both cases, objects under comparison are first compared by a symmetric pairwise matrix.

In gene expression analysis, Euclidean metric distance and Pearson correlation coefficient are the

two most commonly used for comparisons. [176]. However, due to the high-dimensional nature of expression data, both Euclidean distance and Pearson correlation-based norm have significant limitations in reflecting relationships between expression data. For example, Euclidean distance matric is sensitive to scaling. On the other hand, Pearson correlation coefficients only indicate linear relationships between objects and may produce misleading clustering results if outliers are present or if the data distribution is non-Gaussian [177, 178]. These issues have motivated and informed the development of model-based clustering. In model-based algorithms, data are considered as coming from a prior distribution or a mixture of probability distributions, with each representing a distinct cluster [179]. Compared to pair-wise agglomerative (bottomup) classical hierarchical clustering methods, model-based clustering methods provide several advantages. For example, they can use a probability-based approach to determine the optimal number of clusters and no pre-clustering assumption is required [180]. In addition, model-based clustering is less sensitive to outliners and data scaling [181].

In this chapter, bayesian hierarchical clustering (BHC), a model-based clustering algorithm based on the *Dirichlet* process mixture distribution has been used to stratify breast cancer patients. BHC assumes the data is naturally organised as an outcome of a hidden branching process (see Appendix A). Due to this, BHC is expected to reveal a natural stratification of the input objects (here breast cancer patients) according to the distribution of the variables (gene expression variations). Moreover, it is non-parametric and can produce two-dimensional descriptions, dendrogram for both x and y axis, revealing the latent structure in the data.

BHC uses a *Dirichlet* Process (infinite mixture model) to model the uncertainty in the data and Bayesian model selection to determine the hierarchical structure [182, 183]. This allows BHCbased clustering to overcome several key limitations of the classical clustering methods, such as determining optimal clustering number and appropriate distance matrix. This clustering has been implemented for R Bioconductor, widely used and shown to discriminate more biologically meaningful (more biologically homogeneous) clusters than agglomerative hierarchical clustering [184]. Moreover, it has been demonstrated, in a comparative study, that multivariate Gaussian based clustering approach exhibited the best performance in revealing the true structure of the gene expression data [185]. Using BHC, Sirinukunwattana *et al.* tested BHC algorithm over 11 cancer datasets and showed that the partitions revealed are more concordant with the ground truth compared to other classical hierarchical clusterings [186]. Taken together, these studies suggest that BHC is a robust clustering algorithm for analysing gene expression data than other traditional clustering approaches and may provide more informative (biologically relevant) clusters. In this work, BHC was used to analyse two breast cancer gene expression datasets with the hope of revealing novel stratification between subtypes.

#### 2.1.3 Network modelling

Breast cancer, like other cancers, is recognised as a heterogeneous disease containing diverse cell types and tissue origins. It has also been demonstrated that breast cancer oncogenesis involves multiple aberrant signalling events that regulate essential cell processes such as death, proliferation, differentiation and migration [187, 188]. Hornberg *et al.* has argued that tumour development as a whole is expressed as concerted and coordinated reactions of intracellular networks and, due to this, cancer can be referred to as a "systems biology" disease [189]. These researchers demonstrated that biological systems are complex, and the phenotypic diversity is a reflection of a combination of subtle dynamic interactions rather than individual genetic alterations (e.g. mutations, transcriptions, regulation events). It is important to consider breast cancer oncogenesis in the context of a background of dynamic processes and molecular interactions at the system level.

Network modelling provides a useful approach for formally describing and visualising biological systems [190, 191]. A network is a formal description of a collection of object nodes (A), each connected to processes (B). If only nodes (A) are projected, then their multiple connections are edges representing the processes. The partite organisation reflects the objects and their processes (B). In biological networks, a projection is used, nodes represent various biological entities and edges represent inter-relations or reactions between them. Common biological networks include graphs of protein-protein interaction (PPI) networks, signalling networks, gene regulatory networks or, as described here, gene correlation networks (GCNs). A GCN is a signed graph with each node representing a gene and each edge representing a direction as a positive or a negative correlation between a pair of genes [192]. These are co-expression gene relationships, and the correlations reveal genes that are functionally related, controlled by the same set of transcriptional factors or involved in the same pathway or protein complex [193]. Unlike gene regulatory networks, where the relationships between the genes are assumed to reflect a biological process (e.g. activation or inhibition), edges in GCNs only reflect dependency relationships and do not reveal causal relationships between genes [194, 195]. Nevertheless,

co-expression networks are useful in identifying biomolecules that show a unique coordinated expression pattern across a group of samples. Using GCNs, Oh *et al.* have captured critical mRNA relationships in a cohort of breast cancer patients and identified an important crosstalk event in breast cancer [196]. GCNs have also been successfully used to identify network-based prognostic modules in cancers [197, 198]. These researchers demonstrated the ability of GCNs in modelling the cellular signalling dynamics and capturing the underlying molecular state driving oncogenesis.

In this work, gene correlation networks were used to describe transcriptional relationships in Basal-like breast cancer. From the correlation networks, three hub-associated network modules (annotated as motifs) were found to be enriched in Basal-specific networks. Graphically, these network motifs are topologically central connected to a hub nuclear receptor (with the maximal number of connections), representing critical positions in the network of nuclear receptors for targeting. In terms of gene transcription, these critical modules are network-based indicators of essential components underlying Basal-like transcriptional regulatory signalling, providing a mechanistic understanding of Basal-like breast cancer characteristics.

# 2.1.4 Aims

This chapter adresses the following two areas:

- Revealing heterogeneity within and between breast cancer subtypes.
- Identifying nuclear receptor functional units that are implicated in Basal-like subtype.

# 2.1.5 Chapter overview

The strategy presented in this chapter is shown in Figure 2.1. The method has three main steps, including data collection and preparation, breast cancer stratification from unsupervised clustering and network modelling and analysis. Pictures on the left show a typical outcome from each step.



Figure 2.1: **Overview of the approach presented in this chapter.** This approach contains three main steps, including data collection and preparation, breast cancer stratification by BHC followed by network modelling and analysis. Pictures on the left show a typical outcome from each step.

Data collection process and methods used in this chapter are described in Section 2.2. patient clustering results produced by BHC are described in Section 2.3. The process of identifying critical NR-based functional network units implicated in Basal-like breast cancer is demonstrated in Section 2.4 and Section 2.5. Limitations and chapter discussion, including future perspectives, are described at the end of the chapter in Section 2.6.

# 2.2 Clustering materials and methods

In this work, nuclear receptor and NR-associated gene expression data for individual patients were analysed using a computational pipeline written in R. Code used for clustering analysis is provided in Appendix A. Data analysis required R (v3.5.3), R Studio(v1.2.5019), R/Bioconductor (3.10), R/BHC (v1.38.0), R/GeneNet 143 (1.2.13) and R/Graphiz (v2.26.0). Two breast cancer cohorts, the cancer genome atlas (TCGA) and molecular taxonomy of breast cancer international consortium (METABRIC), were considered and analysed in this chapter and both datasets were accessed from cBioPortal (https://www.cbioportal.org).

## 2.2.1 TCGA expression data

TCGA (The Cancer Genome Atlas Program) is a US government-funded project that aimed to use high-throughput genome sequencing techniques to characterise genetic variations underlying cancers (https://www.cancer.gov/tcga). A pilot study was initially conducted to evaluate the feasibility of the full-scale effort to systematically explore the wide spectrum of genomic changes involved in human cancer. TCGA has successfully generated and analysed more than 11,000 individuals with more than 30 types of cancers. More than 2.5 petabytes of data were generated from the project. TCGA is the largest published RNA-seq-based breast cancer dataset to date. In addition to mRNA expression data, TCGA also collected a wide rane of genomic data (e.g. copy number, DNA, microsatellite instability, miRNA) as well as clinical information. Data generated from this project is publicly available on cBioportal for researchers in the community to study and validate further.

TCGA project recruited a total number of 825 breast cancer tumour samples and among them mRNA expression data were available for 526 individuals [199]. Importantly, in TCGA, each tumour was assigned to one of the four PAM50 subtypes (Luminal A, Luminal B, Her2 and Basal-like) and in the following analysis, patient subtypes refer to this PAM50 subtype assignment provided by TCGA. Due to insufficient information (missing data), 12 individuals were eliminated from the study, leaving a total number of 514 breast tumour samples. Out of the 514 tumour samples 231 were Luminal A samples, 127 were Luminal B samples, 58 were Her2 samples and 98 were Basal-like samples (Table 2.1).

The mRNA gene expression data from the TCGA study is organised into levels and level 3 is used in this study. Level 3 is the least processed, where the expression signals are median-

centred to account for the potentially skewered distributions of the differentially expressed genes. A total number of 178 genes were considered in this work, including 48 human NRs and their coregulators. The NR coregulators are defined as the genes that code for the functional partners of each NRs. The coregulators are described in the STRING database (v11.0). The complete list of NRs and their corresponding coregulators are described in Appendix B. Out of the 178 NR-associated genes, 171 were available in TCGA and included in the subsequence analysis (data availability is described in Appendix C).

# 2.2.2 METABRIC expression data

Different from TCGA, METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) is only focused on breast cancer. It is a UK-Canada breast cancer genomic project that integrated genomic and transcriptomic profiles from a large breast cancer cohort. A total number of 2509 primary breast tumours were sequenced in this study. It is the largest breast cancer dataset available on cBioPortal [200] to date. The study revealed that there are at least 11 underlying breast cancer subgroups. More recently in 2019, the METABRIC team compared these subgroups' molecular profiles with long-term clinical information, and suggest that these subgroups are clinically distinct, with some groups facing higher risk of risk of recurrence. The genome data as well as the clinical data from METABRIC are also available on cBioportal [201].

METRABRIC project has sequenced 679 Luminal A samples, 461 Luminal B samples, 176 Her2 samples and 199 Basal-like samples. Similarly, each of the tumour samples in METABRIC was also assigned one of the four PAM50 subtypes (Luminal A, Luminal B, Her2 and Basal-like), and this subtype assignment was used to assume samples' molecular subtype in the subsequence analysis. Nuclear receptor and NR-associated gene expression was median-centred using R to account for the potentially skewered distributions of the differentially expressed genes. Out of the 178 NR-associated genes, 169 were available in METABRIC and included in the METABRIC analyses (data availability is described in Appendix C).

Table 2.1: Sample subtype distribution in TCGA and METABRIC.

	Luminal A	Luminal B	Her2	Basal-like
TCGA	231	127	58	98
METABRIC	679	461	220	199

# 2.2.3 Unsupervised Bayesian Hierarchical Clustering

The clustering analysis was performed in R (v3.5.3) and R Studio(v1.2.5019) using R/BHC (v1.38.0) from R/Bioconductor (3.10). The algorithm and concept of Bayesian Hierarchical Clustering is described in detail in Appendix A. Code used for the R/BHC clustering is also provided in Appendix A, to reproduce the results.

BHC is expected to reveal a natural stratification of the input objects (breast cancer patients) according to the distribution of the variables. Patients with the most similar gene expression patterns were expected to group together and those with sufficient dissimilarities were expected to separate into different patient groups. In this work, patients were discriminated into clusters by BHC. The number of patients in a given cluster can be small, sometimes represented by outliers. Therefore, to simplify the patient groupings, the clusters were enumerated into classes, defined according to the merges further up the dendrogram. The original clusters revealed by BHC and the clusters-to-classes simplification process is provided and illustrated in Appendix D. It is important to note that, although classes are a simplification of the clusters, the choice of classes still produces distinct and significantly different groups with distinct NR expression patterns.

In this work, the pairwise comparisons between subtypes provide identification of unique NR expression patterns that are solely implicated in Basal-like patients. When two subtypes are compared, a patient class revealed by BHC is expected to be either dominated by one of the two subtypes or not dominated by either of the subtypes (Figure 2.2). Classes were defined as either a "subtype dominant" or "ambiguous". If a class contains patients predominantly from one subclass, then it is defined as a "subtype dominant". On the other hand, if a class contains a similar number of patients from both subtypes, then the patient class is defined as an "ambiguous".



Figure 2.2: Schematic illustration of subtype dominant class and ambiguous class. In a pairwise comparison (between two subtypes) clustering analysis, resulting clusters are expected to contain predominantly one subtype patients or a mixture of both subtypes. If a class is dominated by one subtype, then it is defined as a "subtype dominant class". If a class is not dominant by neither of the breast cancer subtypes, then it is defined as an "ambiguous class".

#### 2.2.4 Partial correlation networks

Partial correlation networks were derived from the Basal-like dominant classes using the R/-GeneNet library. GeneNet is available from the CRAN repository and from the webpage http://www.strimmerlab.org/software/genenet/. This package is specifically designed for analysing high-dimensional data obtained from high-throughput assays, such as expression microarrays or metabolic profiling. The partial correlation networks produced by R/GeneNet are graphical gaussian models (GGMs) that represent multivariate dependencies in biomolecular networks by means of partial correlation. The R/GeneNet offers two shrinkage estimator options, and in this research the shrinkage estimator implemented in the R/corpcor package was chosen to calculate a reliable estimation of partial correlation matrix. The other shrinkage estimator employs the function R/longitudinal:dyn.pcor from the longitudinal package. The shrinkage estimator described in the longitudinal package is intended for analysing time series data, therefore not applicable for this research. The details of the two shrinkage estimators used in R/GeneNet are described by Opgen-Rhein and Strimmer [202] and Schafer and Strimmer [203]. The function network.test.edges in R/GeneNet was used to create a pairwise partial correlation matrix containing all possible correlations, listed by order of magnitude. An empirical marginalised thresholding of 1% was chosen to reveal strong correlations and eliminate the weaker correlations, after testing a series of threshold levels. The choice of thresholding is discussed later in Section 2.6.2. R/Graphiz was used to generate networks for visualisation. Each node in the network represents a gene and the lines between them represent a strong (top 1%) correlation. Each nodes were marked with degrees, defined as the number of connections that a given node makes in a network. The node degrees were obtained from R/node.degree. Total degree is the sum of degree each node makes across the eight Basal-specific networks. All nodes are ranked by their total degree. The total degree distribution of the nodes is also plotted for hub identification. Hubs are the top three nodes with highest degree (number of connections). The networks developed in this work are signed (containing positive and negative correlations) and undirected.

## 2.2.5 Hub-associated local networks and 3-node motifs

Once the hubs are identified, the hub-associated local network defined by 3-node motifs can be identified. Hub-associated local networks are defined by the inclusion of all possible 3-node motifs within the proximity of two edges from a hub node. In this work, 3-node motifs centre around the top two most connected nodes were annotated and compared for all networks derived from Basal dominant classes. Motifs and hub-associated local networks were curated manually and checked systematically to eliminate human error in the annotation process.

# 2.3 Stratification comparison by BHC

In this chapter, breast cancer patients were stratified into natural groups by BHC according to their nuclear receptor and NR-associated gene expression levels. To reveal unique transcriptional characteristics implicated in BLBC, it was compared to non-BLBC subtypes and three pairwise comparisons performed:

- Basal-like vs Luminal A
- Basal-like vs Luminal B
- Basal-like vs Her2

Bayesian Hierarchical Clustering of patient-gene expression data can be performed in two directions, giving rise to two clustering interpretations. Based on different marginal likelihood estimations, the result gives two distinct discretised colour plots with the marginal likelihood (in black) and the silent contributions lower bounds (in green) and upper bounds (in red), with the resulting dendrograms on alternative axes. One clustering result uses the transform of the dataset to the other. In one result, patients are stratified according to the information content from the distribution of the individual gene expression distributions. In the other dimension, nuclear receptor and associated genes are clustered according to information content from the distribution of the individual patient profiles, which are non parametric. The two sets of clustering interpretations are asking different questions about the data: (i) the patient clustering analyses are expecte to reveal heterogeneous transcriptional patterns in different BC subtypes, and (ii) the nuclear receptor gene clustering analyses are expected to reveal group(s) of nuclear receptor or associated genes that are unique (e.g. highly expressed) to Basal-like breast cancer. In this chapter, the patient clustering is shown. The complementary analysis, nuclear receptor gene clustering, is described and discussed in the following Chapter 3.

In the following cluster analyses, a corresponding "2D colour plot" is aligned with the 1D dendrograms at the top, showing the hierarchy of the patient clusters. For each gene, the continuous expression values were discretised into three groups (the upper-bound representing high-expression, the marginal-likelihood represents average behaviour, and the lower-bound representing low-expression) prior to clustering. Effectively, the high-expression genes are represented by red, low-expression genes are represented by green. Importantly, both do not contribute to the clustering and are referred to as silent contributions. In the following figures, yellow lines show the organisation of classes (simplified from the clusters) and the corresponding expression signal pattern in the colour plot.

## 2.3.1 Basal-like vs Luminal A

In the Basal-like vs Luminal A analysis, a total number of 329 patients (98 Basal-like 262 and 231 Luminal A) from TCGA database were grouped into five classes (Figure 2.3, Left). Similarly, 878 patients (199 Basal-like and 679 Luminal A) from METABRIC were grouped into six classes (Figure 2.3, Right). The classes are a simplification of clusters revealed by BHC, representing patient groups with distinct NR transcriptional patterns. In both analyses, class one is considered as Basal dominant classes, as they consist mostly of Basal-like patients

with only one Luminal A patient in TCGA class 1. Class three, four and five from both TCGA analysis and METABRIC analysis are considered as Luminal A dominant classes as the majority of the patients fell into these classes are Luminal A patients. Class two from TCGA analysis and class two and six from METABRIC analysis are considered as ambiguous classes as they consist a mixture of Basal-like and Luminal A patients. The hierarchical clustering structure (shown in the top dendrograms) shows that Basal dominant classes from both analyses are separated from the Luminal A dominant classes at the top divisions, suggesting that the patients falling into Basal dominant classes have very distinct nuclear receptor signal patterns comparing to those fall into Luminal A dominant classes.



Figure 2.3: Extracted BHC results from Basal-like vs Luminal A comparison shows a natural stratification of Basal-like vs Luminal A patients (columns) according to the NR-associated gene expression signals (rows). Left: Patients from TCGA dataset were grouped into five classes. Class 1 is considered as a Basal dominant class. Right: Patients from METABRIC dataset were grouped into six classes. Class 1 is considered as a Basal dominant class. Yellow lines show the organisation of classes and the corresponding signal pattern in the colour plot. The tables show the number of Basal-like and Luminal A patients in each class. The Basal dominant classes are highlighted in red boxes in the table and colour plots.

# 2.3.2 Basal-like vs Luminal B

Similar results were obtained in the Basal-like vs Luminal B comparisons. Patients were grouped into three and seven classes in the TCGA and METABRIC analyses, respectively (Figure 2.4). Class 1 from the TCGA analysis and class 1 and class 2 from the METABRIC analysis were considered as Basal dominant classes. While two Basal dominant classes (class 1 and class 2) were identified from the METABRIC, the class 1 in METABRIC is assumed to present less Basal-like characteristics than class 2 as it only contains 50 Basal-like patients and 13 Luminal B patients whereas class 2 contains 111 Basal-like patients and 3 Luminal B patients. Thus, METABRIC class 1 is considered as an insubstantial Basal dominant class. In both TCGA and METABRIC analyses, a clear separation was observed at the top division of the hierarchy, separating the Basal dominant classes from Luminal B patients have distinctive NR expression signal patterns.



Figure 2.4: Extracted BHC results from Basal-like vs Luminal B comparison shows a natural stratification of Basal-like vs Luminal B patients (columns) according to the NR-associated gene expression signals (rows). Left: Patients from TCGA dataset were grouped into three classes. Class 1 is considered as a Basal dominant class. Right: Patients from METABRIC dataset were grouped into seven classes in which class 1 and class 2 are considered as Basal dominant classes. Yellow lines show the organisation of classes and the corresponding signal pattern in the colour plots. The tables show the number of Basal-like and Luminal B patients in each class. The Basal dominant classes are highlighted in red boxes in the table and colour plots.

# 2.3.3 Basal-like vs Her2

Unlike the Basal-like vs Luminal type comparisons shown in the previous sections, the separations between Basal-like and Her2 are not as clear in Basal-like vs Her2 comparisons. In the TCGA analysis (Figure 2.5, Left), Basal-like and Her2 patients were organised into four classes with two of them being Basal dominant classes (class 1 and class 3). Interestingly, the two Basal dominant classes fell into different descenders at the top division of the hierarchical structure, suggesting that the two Basal dominant classes have distinct NR expression patterns. Class 1 fell into the left descender and is separated from the rest of the classes, whereas class 3 fell into the right descender with the Her2 dominant class (class 2) and the ambiguous class (class 4). TCGA class 3 shares more similarity with Her2 therefore is considered as an insubstantial Basal dominant class.

In the METABRIC analysis (Figure 2.5. Right), patients were grouped into seven classes. While more than half of the Basal-like patients (115 out of 199 Basal-like patients) fell into the Basal dominant class (class 2), the rest of the Basal-like patients were distributed across the other six classes. In particular, 46 out of 199 Basal-like patients fell into the Her2 dominant class (class 3), suggesting that these patients have similar NR expression patterns to Her2 patients.



Figure 2.5: Extracted BHC results from Basal-like vs Her2 comparison show a natural stratification of Basal-like vs Her2 patients (columns) according to the NRassociated gene expression signals (rows). Left: Patients from TCGA dataset were grouped into four classes. Class 1 and Class 3 are considered as Basal dominant classes. Right: Patients from METABRIC dataset were grouped into seven classes. Class 2 is considered as a Basal dominant class. Yellow lines show the organisation of classes (simplified from the clusters) and the corresponding signal pattern in the colour plot. The tables show the number of Basal-like and Her2 patients in each class. The Basal dominant classes are highlighted in red boxes in the table and colour plots.

The BHC analyses described above show that while the majority of the BLBC patients are separated from Luminal type patients at the top division of the dendrograms, a fraction (nearly a half) of Basal-like patients from both TCGA and METABRIC were shown to have more similar NR expression patterns to Her2 patients than to other BLBC. This observation is consistent in both TCGA and METABRIC analyses.

# 2.4 Basal-specific correlation networks

To further explore the NR expression patterns underlying BLBC, pairwise partial correlation between NR-associated genes were calculated to examine the NR associations amongst Basal-like patients. Partial correlation measures the strength of linear relationships (associations) between two variables (e.g. gene expressions). Different from regular correlation, partial correlation eliminates the effect of other variables. This is particularly important and useful for genomic studies because it allows detection of conditional independence between two variables [204, 205]. The partial correlation approach has been successfully applied to reveal and understand the complexity of transcriptomic data in yeast [206, 207], HeLa cells [208], and breast cancer tumours [209].

In this work, partial correlation-based networks were constructed to demonstrate the dependency associations between NR-associated genes in Basal-like patients. Networks were derived from different Basal dominant classes revealed by BHC (described in Section 2.3.1 to 2.3.3), referred to as Basal-specific networks. Each node in the Basal-specific networks represents a gene, and each edge representing an association (quantified by the the partial correlation) between two genes. Since correlation is symmetrical, the networks developed in this work are signed (containing both positive and negative correlations) and undirected. Positive correlations are shown in red and negative correlations are shown in black. Figure 2.6 shows an example of the partial correlation networks developed from Basal dominant class from TCGA Basal-like vs Luminal A comparison. In this work, 8 Basal-specific networks were developed, and the networks derived from the other comparisons are shown in Appendix E.



Figure 2.6: An example of Basal-specific partial correlation networks constructed from Basal dominant class TCGA class 1 from the Basal-like vs Luminal A comparison. Each node represents a gene and the lines between the nodes represent a correlation. Red lines represent positive correlations and black lines represent negative correlations. The numbers next to the lines represent the strength of the partial correlations. Only the strongest 1% correlations were included in the networks.

# 2.5 Motif classifications and identification

To identify critical nuclear receptor and NR-associated genes across the 8 Basal-specific networks, a frequency analysis was performed to summarise the *degree* of each node in all eight Basal-specific networks. The *degree* of a node is the number of associations that are connected to that node. The *total degree* is the sum of individual *degree* each node makes across the
eight Basal-specific networks. All nodes were ranked according to the magnitude of *total degree* across the networks. The distribution of the *total degree* of the nodes was plotted and is shown in Figure 2.7. fos proto-Oncogene (FOS) and STAT1 have typically demonstrated the highest number of connections of any of the nodes measured within the eight Basal-specific networks. The table in Figure 2.7 shows the top 15 most connected nodes across networks, with FOS having a *total degree* of 28 connections and STAT1 having a *total degree* of 26 connections. FOS and STAT1 are the two most connected nodes across networks and are defined as hubs. Hub genes have the highest *degree* of connectivity and therefore occupy central and critical positions in all networks, suggesting a more substantial capacity to modulate adjacent genes than the genes with the lower degrees. The complete summary table containing all the 178 nodes and their corresponding degrees across networks is shown in Appendix F.

In order to reveal critical modules embedded in Basal-specific correlation networks, associations around the hubs were explored. Local networks centred around the hubs defined by 3-node motifs were extracted from each Basal-specific network and are shown in Figure 2.8 to Figure 2.11. Hub-centred local networks contain up to two levels of correlations from the hubs. They were used to identify critical 3-node motifs that are self-consistent across the network analyses. Three-node motifs are the smallest functional units in a network and are the fundamental building block for lager motifs (4-node or 5-node motifs). They are often a useful indicator for revealing functional properties in a network [210].



В

NR	Basal vs Luminal A		Basal vs Luminal B				Total		
Identifier	TCGA Class1	METABRIC Class1	TCGA Class1	METABRIC Class1	METABRIC Class2	TCGA Class1	TCGA Class3	METABRIC Class2	Total
FOS	3	4	3	3	5	5	1	4	28
STAT1	4	4	2	3	4	3	3	3	26
NR2C1	3	2	3	5	3	2	4	2	24
NR4A2	4	4	3	0	3	3	3	4	24
CREBBP	3	5	4	1	4	2	0	4	23
NCOR2	3	3	3	3	3	2	3	2	22
APOA1	3	2	4	1	1	3	4	3	21
LCK	4	3	3	1	3	2	3	2	21
TMPRSS2	3	4	3	3	2	2	1	3	21
POMC	1	3	1	4	2	3	4	2	20
SP1	3	1	5	3	2	1	4	1	20
THRSP	3	3	2	2	1	2	5	2	20
USF2	3	1	2	4	0	2	5	3	20
COPS2	0	4	2	3	4	2	2	2	19
HNF4A	3	3	2	2	3	2	1	3	19

Figure 2.7: Node *degree* summary. (A) The histogram shows the distribution of *total degree* of the nodes. FOS and STAT1 are the most connected nodes across networks with a *total degree* of 28 and 26, respectively. (B) The 15 most connected genes listed by the total number of connections in all Basal-specific networks. While FOS and STAT1 have typically demonstrated the highest number of connections of any of the nodes measured within the eight "Basal-specific networks". It should be considered though that there are variances amongst the networks, with FOS ranging from 1-5 and STAT1 ranging from 3-5 connections per network.



Figure 2.8: **FOS-centred local networks from TCGA analyses.** FOS is coloured in yellow and its neighbouring genes are coloured in blue. Positive edges (correlations) are in red and negative edges are in black.



Figure 2.9: **FOS-centred local networks from METABRIC analyses.** FOS is coloured in yellow and its neighbouring genes are coloured in blue. Positive edges (correlations) are in red and negative edges are in black.



Figure 2.10: **STAT1-centred local networks from TCGA analyses.** STAT1 is coloured in yellow and its neighbouring genes are coloured in blue. Positive edges (correlations) are in red and negative edges are in black.



Figure 2.11: **STAT1-centred local networks from METABRIC analyses.** STAT1 is coloured in yellow and its neighbouring genes are coloured in blue. Positive edges (correlations) are in red and negative edges are in black.

In an undirected and signed correlation network, there are 27 possible configurations of a 3node motif (Figure 2.12). They are non-redundant motifs, if each node is defined clockwise labelling (e.g. A, B, C), they can be classified into 10 NPU (Negative, Positive, Un-associated) groups, according to a classification scheme adapted from the concept of triad census proposed by Davis and Leinhardt [211]. The classification scheme assigns each motif an NPU code, where the first number represents the number of Negative correlations (black); the second number being the number of Positive correlations (red) and the third number represents the number of Un-associated correlations in a given 3-node motif. For example, a 003 motif would have three un-associated edges seemingly independent. Likewise, a 300 motif would have three negative correlations, and a 030 motif would have three positive correlations. Within each NPU type, 3-node motifs may have different variations; for example, motif type 111 has six different configurations. Type 021, 201 and 111 contain three nodes and two edges, forming linear topology. Type 120, 210, 030, 300 all contain three nodes and three edges, forming complete topology. If each node is given clockwise labelling (e.g. A, B, C), there are 12 and 8 possible configurations of linear and complete 3-node network motifs from an undirected correlation network, respectively. The linear and complete topology are explained in Figure 2.13.

NPU	3-node n	notifs config	gurations	# of edges	# of configurations
003		8 © 8		0	1
012	Ø Ø	© ®	8 ©®	1	3
102	Ø 0	© ®	0 ©0	1	3
021	o B	Ø 6 8	© ®	2	3
201	o Po	©B	©B	2	3
111				2	6
120				3	3
210				3	3
030		C B		3	1
300				3	1

Figure 2.12: NPU classification of 3-node motifs. Nodes are labelled A, B, C clockwise to represent non-redundant NRs. Edges are illustrated as positive correlations (stronger than +1.0%) in red or as negative correlations (stronger than -1.0%) in black. The three-number NPU code represents the number of negative correlations, positive correlations and un-associated correlations in a given motif, respectively. The classification scheme was adopted from the concept of the triad census proposed by Davis and Leinhardt [211].



Figure 2.13: Linear and complete motifs. (A) The 12 configurations of the linear motifs based on the position of the hub (labelled A and in yellow). Top row: node A is connected to both B and C. Middle row: node A is connected to C via B. Bottom row: node A is connected to B via C. (B) The eight configurations of the non-redundant complete 3-node network motifs. These motifs are divided into two groups based on their characteristics (coherence). The hubs (node A) are shown in yellow. Top row: Node A and associated nodes B and C and are understood to have coherent behaviour. Bottom row: Node A and associated nodes B and C and are understood to have incoherent behaviour.

The occurrence of the linear (021, 201, 111) and complete (120, 210, 030, 300) motifs around the hubs in each local network was summarised in Table 2.2. In most networks, motif type 021 is the most common form of 3 node motifs followed by type 111 and then type 030.

Table 2.2: Occurrence of the seven types of NPU motifs centred around FOS and STAT1 from Basal-specific networks. Numbers in the table indicate the frequency of an NPU motif occurs in hub-associated local networks.

FOS		021	201	111	120	<b>210</b>	030	300
Basal-like vs Luminal A	TCGA Class1	8	0	0	0	0	2	0
	METABRIC Class 1	8	0	0	0	0	2	0
Basal-like vs Luminal B	TCGA Class1	7	0	1	0	0	0	0
	METABRIC Class 1	4	0	2	0	0	0	0
	METABRIC Class 2	8	0	6	0	0	1	0
Basal-like vs Her2	TCGA Class1	14	0	2	0	0	0	0
	TCGA Class3	0	0	0	0	0	0	0
	METABRIC Class 2	9	0	1	0	0	2	0
					100			
STAT1		021	201	111	120	210	030	300
STAT1 Basal-like vs Luminal A	TCGA Class1	<b>021</b> 9	<b>201</b> 0	<b>111</b> 0	<b>120</b> 0	<b>210</b> 0	<b>030</b>	<b>300</b> 0
STAT1 Basal-like vs Luminal A	TCGA Class1 METABRIC Class 1	<b>021</b> 9 8	<b>201</b> 0 0	<b>111</b> 0 1	<b>120</b> 0 0	<b>210</b> 0 0	<b>030</b> 0 1	<b>300</b> 0 0
STAT1 Basal-like vs Luminal A Basal-like vs Luminal B	TCGA Class1 METABRIC Class 1 TCGA Class1	<b>021</b> 9 8 2	<b>201</b> 0 0 0	<b>111</b> 0 1 0	<b>120</b> 0 0 0	<b>210</b> 0 0 0	<b>030</b> 0 1 0	<b>300</b> 0 0 0
STAT1 Basal-like vs Luminal A Basal-like vs Luminal B	TCGA Class1 METABRIC Class 1 TCGA Class1 METABRIC Class 1	021 9 8 2 2	<b>201</b> 0 0 0 0	<b>111</b> 0 1 0 3	<b>120</b> 0 0 0 0	<b>210</b> 0 0 0 0	030 0 1 0 0	<b>300</b> 0 0 0
STAT1 Basal-like vs Luminal A Basal-like vs Luminal B	TCGA Class1 METABRIC Class 1 TCGA Class1 METABRIC Class 1 METABRIC Class 2	021 9 8 2 2 8	<b>201</b> 0 0 0 0 0	<b>111</b> 0 1 0 3 0	<b>120</b> 0 0 0 0 0 0 0 0	<b>210</b> 0 0 0 0 0	030 0 1 0 0 1	<b>300</b> 0 0 0 0 0
STAT1 Basal-like vs Luminal A Basal-like vs Luminal B Basal-like vs Her2	TCGA Class1 METABRIC Class 1 TCGA Class1 METABRIC Class 1 METABRIC Class 2 TCGA Class1	021 9 8 2 2 2 8 6	201 0 0 0 0 0 0 0	111 0 1 0 3 0 0	<b>120</b> 0 0 0 0 0 0 0 0 0 0 0 0	210 0 0 0 0 0 0 0	030 0 1 0 0 1 1 1	<b>300</b> 0 0 0 0 0 0
STAT1 Basal-like vs Luminal A Basal-like vs Luminal B Basal-like vs Her2	TCGA Class1 METABRIC Class 1 TCGA Class1 METABRIC Class 1 METABRIC Class 2 TCGA Class1 TCGA Class3	021 9 8 2 2 8 8 6 4	201 0 0 0 0 0 0 0 0	111 0 1 0 3 0 0 0 0	120 0 0 0 0 0 0 0 0	210 0 0 0 0 0 0 0 0	030 0 1 0 0 1 1 1 1	<b>300</b> 0 0 0 0 0 0 0 0

Three 3-node motifs jun Proto-Oncogene (JUN)-FOS-NR4A2, promyelocytic leukemia protein (PML)-STAT1-GCH1 and STAT1-LCK-NR1H3 were found to be enriched in the Basalspecific networks from both TCGA and METABRIC datasets (Figure 2.14 to Figure 2.16). Both motifs, JUN-FOS-NR4A2 and PML-STAT1-GCH1, formed a 021 linear topology and were found in six out of eight networks (Table 2.3). The motifs were absent from METABRIC class 1 from Basal-like vs Luminal B comparison and TCGA class 2 from Basal-like vs Her2 comparison. Both classes were classified as insubstantial Basal dominant classes from the corresponding comparisons (see Section 2.3.2 to 2.3.3). Motif STAT1-LCK-NR1H3 forms a positive linear topology (type 021) and positive complete topology (type 030) in two and three Basalspecific networks, respectively. The 030 coherent (positive) complete motifs represent positive coordination between the three entities (genes), suggesting a tight positive transcriptional regulation. The same set of genes but with a slightly different arrangement, STAT1, LCK and NR1H3 also form a STAT1-NR1H3-LCK 021 motif in the METABRIC class 1 from Basal-like

#### vs Luminal B comparison.



Figure 2.14: Motif JUN-FOS-NR4A2 embedded in the six out of eight Basal-specific networks. The 3-node motifs are highlighted in bold in each network, with the hubs in yellow and non-hub nodes in blue. The numbers of the edges are the strength of the corresponding correlation.



Figure 2.15: Motif PML-STAT1-GCH1 embedded in the six out of eight Basalspecific networks. The 3-node motifs are highlighted in bold in each network, with the hubs in yellow and non-hub nodes in blue. The numbers of the edges are the strength of the corresponding correlation.



Figure 2.16: Motif STAT1-LCK-NR1H3 embedded in the six out of eight Basalspecific networks. The 3-node motifs are highlighted in bold in each network, with the hubs in yellow and non-hub nodes in blue. The numbers of the edges are the strength of the corresponding correlation.

Table 2.3: Occurrence of the critical 3-node motifs enriched in Basal-specific networks. Network motif JUN-FOS-NR4A2 and PML-STAT1-GCH1 both form type 021 linear motifs in the networks, whereas network motif STAT1-LCK-NR1H3 form both type 021 linear motifs and type 030 complete motifs. Motif type is described in the bracket.

	Basal-like vs Luminal A		Basal-like vs Luminal B			Basal-like vs Her2		
	TCGA class 1	METABRIC class 1	TCGA class1	METABRIC class 1	METABRIC class 2	TCGA class 1	TCGA class 3	METABRIC class 2
IUN FOS ND442	Present	Present	Present	Absent	Present	Present	Absent	Present
JUN-F05-M04A2	(linear, 021)	(linear,021)	(linear,021)		(linear, 021)	(linear, 021)		(linear, 021)
DML STATLCCUL	Present	Present	Present	Abcont	Present	Present	Abcont	Present
I ML-SIAII-GOIII	(linear, 021)	(linear, 021)	(linear, 021)	Absent	(linear, 021)	(linear, 021)	Absent	(linear, 021)
STAT1 LOW ND 1112	Present	Present	Abcomt	STAT1-NR1H3-LCK	Present	Abcomt	Present	Present
51A11-LUK-INKIN5	(linear, 021)	(complete, 030)	Absent	(linear, 021)	(complete, 030)	Absent	(complete, 030)	(linear, 021)

# 2.6 Discussions and summary

The method presented in this chapter demonstrates a novel approach for analysing and interpreting transcriptomics data. This method integrates a machine learning clustering method (BHC) with network modelling to identify critical NR-based core modules enriched in Basallike expression networks. BHC was used to stratify breast cancer based on their NR expression patterns from two independent breast cancer cohorts (TCGA and METABRIC). The breast cancer subtype heterogeneity revealed by BHC will be discussed in Section 2.6.1 and compared with the current literature. Basal dominant classes were identified according to the stratification revealed by BHC, representing distinct groups of patients with Basal-like characteristics. Partial correlation-based networks were then calculated for each Basal dominant class to describe and visualise the associations amongst genes. The advantages of using partial correlationbased networks instead of full correlation-based networks are described in Section 2.6.2. While marginalised thresholding of 1% was used to eliminate weak correlations, other thresholding criteria were also explored. The justification for the choice of thresholding is also discussed in Section 2.6.2. Critical three-node motifs enriched in Basal-specific networks were identified and their implications are discussed in Section 2.6.3. Other limitations and future perspectives are discussed in Section 2.6.4.

#### 2.6.1 Heterogeneity in breast cancer subtypes

As BHC does not require any pre-clustering assumptions, it uses a probabilistic approach to model the hierarchical structure in big data according to the hidden (latent) groups [186]. It is assumed that BHC reveals a natural and less biased breast cancer stratifications [184].

Based on NR expression patterns, BHC clustering results show a similar composition of tumour subtype to the PAM50 subtype classification [42]. In particular, the majority of Basal-like patients are distinguishable and show distinct NR expression patterns from Luminal A and Luminal B patients. Surprisingly, in both TCGA and METABRIC datasets, a fraction of Basal-like patients (43/89 from TCGA, 46/199 from METABRIC) were shown to share more similarity with Her2 patients than to other Basal-like patients. This observation supports the well-recognised assumption that Basal-like subtype defined by PAM50 subtyping scheme has molecular heterogeneity [212–214]. While the PAM50 classification developed by Parker *et al.* has provided a foundation for breast cancer categorisation, more effort has been put into disambiguating the breast cancer subtypes [213, 215]. For example, Sabatier *et al.* used a 368gene expression signature to classify Basal-like patients into two prognostic subgroups, with one having better clinical outcomes than the other [216]. At the cellular level, Basal-like cell lines are classified into Basal-A and Basal-B, with Basal-A being more Luminal-like and Basal-B more Basal-like [215, 217].

Using a different unsupervised clustering method (BRB-ARRAYTOOLS), Sotiriou *et al.* revealed a similar finding to the research here, where Basal was organised into two natural subgroups (Basal-1 and Basal-2), with Basal-1 showing more similar gene expression signatures to Her2 than to Basal-2 [52]. The Basal heterogeneity observed from Sotiriou *et al.* and this work could indeed reflect different cellular origins, mutations or a combination of the two, and identifying factors that drive the Basal-like diversity is the next challenge [218]. Moreover, while Sotiriou *et al.* revealed distinct Basal-like subgroups analysing the whole genome pattern (of 7,650 probes), in this research, I was able to obtain a similar stratification by using a much smaller and functionally focused subset of genes of NRs and coregulators. By themselves, the NRs and coregulators are important transcriptional regulators for breast cancer oncogenesis and revealing distinct NR expression patterns associated with different subtypes highlights the therapeutic potential of targeting the pathways associated with the crirical NRs and coregulators.

The approach of using pairwise comparisons to compare Basal-like with other PAM50 subtypes allows the identification of similar and distinct characteristics both within and between subtypes, providing the fundamental biological differences underlying heterogeneous breast cancers.

Furthermore, it is important to recognise the presence of ambiguous cases (patients that fall into either an ambiguous class or grouped with other subtypes in the analyses). They represent tumours with a higher degree of heterogeneity and are often more challenging to treat [219]. A major challenge in understanding these ambiguous cases is that the sample size of these cases is often too small to establish systematic analysis that is replicable.

#### 2.6.2 Partial correlation-based network and choice of thresholding

In this chapter, gene correlation networks (GCNs) were used to describe the associations between NR-associated genes. Correlations describe in GCNs often reveal genes that are functionally related (or coherent), controlled by the same set of transcriptional factors or are involved in the same pathway or protein complex [193]. It is important to note that, GCNs are different from gene regulatory networks, where the relationship between genes are assumed to reflect biological processes (such as activation or inhibition). GCNs do not attempt to infer the causality relationship, instead of edges in GCN only reflect dependency relationships.

Gene correlation networks can be constructed using different correlation measures, including full correlations (e.g. Pearson's correlation), partial correlation, regression and mutual information. While full correlation-based networks are popular, previous research has suggested that partial correlation is a more appropriate measure for interpreting the dependencies between variables (e.g. transcripts) in a complex network than full correlation, and partial correlation-based networks may provide more insightful results [205, 220, 221]. This is because full correlation only reflects the correlation coefficients between two given variables (e.g. transcripts), regardless of all the other variables. In contrast, the partial correlation between a given pair of variables is calculated with all the other variables being controlled for. Consequently, partial correlations represent direct associations, whereas canonical correlation (full correlation) analyses do not distinguish between indirect and direct associations. This is particularly useful for inferring transcriptional networks, as partial correlation networks only capture direct associations between variables and limiting the number of spurious (false positives) edges (Figure 2.17).



Figure 2.17: Comparison between canonical correlation network and partial correlation network. (A) A true regulatory network model, where node A regulates node B and node B regulates node C. However, node A does not have a direct effect on node C. (B) Canonical correlation network captures both direct associations (A-B and B-C) and indirect associations (A-C). Indirect associations captured by canonical correlations are referred to as spurious edges. (C) Partial correlation network only captures direct associations and do not reveal indirect correlations as the dependence of other variables is removed. Figure adapted from [222].

Most correlation networks are constructed based on a given threshold to eliminate less signif-

icant associations. While a range of threshold selecting methods has been proposed, the most commonly used method is to choose an appropriate cut-off and to include all correlations above that cut-off in a network. In this network analysis, marginalised thresholding of 1% was used to construct Basal-specific networks. The 1% thresholding was an empirical threshold, assigned after testing a series of different thresholding levels (15%, 10%, 5%, 2%, 1%, 0.5%) to reveal the most informative networks. Loosening the threshold (greater than 1%) increases the network complexity and therefore require automated motif enumeration (e.g. with mfinder [223]). On the other hand, if thresholding is tighter than 1%, the resulting network will be fragmented and not enough edges would be included to reveal any network structure. Basal-specific networks revealed from different thresholding levels are shown in Appendix G. In summary, 1% was the minimal thresholding to reveal a network topology that was easily enumerated and extends beyond the hub nodes.

#### 2.6.3 Implication of 3-node motifs

FOS and STAT1 were defined as hub nodes as they have typically demonstrated the highest number of connections of any of the nodes measured within the eight partial correlation networks derived from Basal dominant classes. They are typically central in the different networks, representing important positions in the networks. To further evaluate the hubs, 3-node motifs were identified around the hubs (Figure 2.14-2.16). Three 3-node motifs (JUN-FOS-NR4A2, PML-STAT1-GCH1 and STAT1-LCK-NR1H3) were consistently found and enriched in the Basal-specific networks derived from TCGA and METABRIC, representing critical network modules in Basal-like breast cancer. The consistency highlights the robustness of the observed patterns, where the partial correlations presented in the motifs can be directly related to mRNAmRNA interactions or protein-protein interactions (PPIs).

The JUN-FOS-NR4A2 network motif was found in most Basal-specific networks apart from the METABRIC class1 from Basal-like vs Luminal A comparison and Her2-like Basal dominant class from TCGA (Basal-like vs Her2 comparison, class3). This motif is linear (021 motif), with FOS in the middle. This suggests that FOS acts as a mediator, critical for the JUN and NR4A2 associations. Sotiriou *et al.* observed a similar breast cancer stratification to our results and reported that Basal-2 (the Basal-like Basal group) showed higher expression of FOS compared to Basal-1 (the Her2-like Basal group) and other breast cancer subtypes [52]. FOS is a proto-oncogene and has been identified as a survival predictor [224] and driver for breast cancer

metastasis formation [225]. At the genetic level, gene FOS, JUN and NR4A2 are all classified as immediate early genes (IEGs) [226]. The IEGs are activated rapidly in response to an external signal and are considered as a primary cellular response to cellular stimuli [227]. In many cancers, including breast cancer, the expression of IEGs has been found to be sustained and overexpressed [228]. This high level of expression is thought to be caused by unchecked, constitutively active mitogen-activated protein kinase (MAPK) signalling pathway that is commonly seen in cancers [229]. At the protein level, gene FOS codes for functional protein c-FOS that dimerises with c-JUN proteins to form the transcription factor complex AP-1 (activator protein 1), promoting breast cancer growth [230]. The dimension of JUN and FOS could indeed explain the strong correlation between JUN and FOS revealed from the partial correlation networks. Mechanically, interfering with the dimer could be one inhibition strategy. Gene NR4A2 codes for nuclear receptor related 1 protein (NURR1) (nuclear receptor related 1 protein), a steroid-thyroid hormone-retinoid receptor. NR4A2 plays an important role in maintaining the dopaminergic system of the brain and mis-regulation of NR4A2 has been associated with a variety of dopaminergic disorders, including Parkinson's disease and schizophrenia [231]. Recent studies have also revealed its oncogenic roles in many cancers, showing its ability to promote or suppress cancer progression depending on the cellular context [232]. In breast cancer, it has been proposed that NR4A2 has a dichotomous role [233]. Although, currently no literature has provided a clear explanation for the FOS-NR4A2 correlation. Both FOS and NR4A2 have been identified as potential drug targets and biomarkers for breast cancer [233–235].

In this work, two network motifs were associated with STAT1, the other hub from the Basalspecific networks. Both STAT1 motifs are made up with genes that are involved in immune response or in associated regulatory events. In one of the STAT1 motifs, STAT1 was shown to correlate with PML (promyelocytic leukemia protein) and GCH1 (GTP cyclohydrolase 1). Gene PML encodes a PML protein, which is an essential component for PML nuclear bodies (a multi-protein sub-nuclear structure). While PML was initially recognised as a tumor-suppressive factor[236], recent studies have demonstrated that the role of PML in breast cancer goes beyond a tumour suppression [237, 238]. In triple negative breast cancers, PML has been shown to be up-regulated and its activities can be crucial for the survival of TNBC cells [239]. Importantly, it has also been evidenced by Hsu *et al.* in preclinical models that PML can positively regulates STAT1/2 isgylation and transcriptional activity [240]. This regulatory relationship observed by Hsu *et al.* could explain the strong positive correlation between the two genes revealed from the network motif. In the same network motif, STAT1 was also associated with GCH1, which is a gene that codes for an enzyme for the biosynthesis of tetrahydrobiopterin. In breast cancer, it has been shown that GCH1 is over-expressed in ER negative breast tumors and this its high level of expression is associated with poor prognosis [241]. In agreement to the STAT1-GCH1 positive correlation observed in the motif, previous studies have also demonstrated that the transcription of the GCH1 gene is mediated by the Jak2–Stat1 pathway [242] and positively regulated by STAT1 [243]. Moreover, in the PML-STAT1-GCH1 motif, all three genes have been shown to have an important role in interferon  $(IFN)\gamma$ -related defence response [244]. This result is in agreement with a recent study done by Thorsson *et al.* where they showed 60.5%Basal-like breast cancer displayed an IFN $\gamma$  immune response signature, whereas only 22.8% of Luminal A, 46.6% of Luminal B, 49.3% of Her2 and 33.3% of normal like patients displayed this signature [245]. IFN $\gamma$  is an interferon produced by T helper cells (specifically, Th1 cells), cytotoxic T cells and macrophages in response to cytokine and antigen stimulation. Results from Thorsson et al. (2018) suggested that the tumour microenvironment of Basal-like breast cancer may compose of a higher proportion of immune cells, cytokines and stroma than other breast cancer subtypes. This could explain the strong associations among the immune-related genes PML, STAT1 and GCH1 in Basal-like samples revealed in this work.

Expression of STAT1 has also been shown to correlate with LCK (lymphocyte-specific protein tyrosine kinase) and NR1H3 (LXR $\alpha$ ) in this work. While the STAT1-LCK and STAT1-NR1H3 correlations have not been reported by previous studies, it is possible that STAT1-LCK and STAT1-NR1H3 correlations represent immune-related cellular communications and cross-talks between signalling pathways. Specifically, mRNA expression of LCK codes for lymphocytespecific protein tyrosine kinase which is a key tyrosine kinase in T cells. Activated T cells produce interferons, including type I interferons (IFN- $\alpha$ , IFN- $\beta$ ). Type I interferons can bind to receptor proteins and activate janus kinase (JAK)/transducer and activator of transcription (STAT) signalling pathway, regulating a series of downstream proteins, including STAT1 activation [246]. Moreover, while NH1H3 is not generally considered to be involved in immune response or associated pathways, results from Pascual-García et al. have suggested that there is a cross-talk between IFN $\gamma$ /STAT1 and LXRs (LXR $\alpha$  and LXR $\beta$ ) [247]. However, whether the STAT1-NR1H3 correlation revealed in this network analysis represents a cross-talk between STAT1 and LXRs in Basal-like tumour needs further evaluation. Importantly, the two STAT1-associated motifs revealed from this study suggested that the tumour microenvironment of Basal-like breast cancer composed of immune cells and understanding the roles of immune system in Basal-like pathogenesis will be the next challenge.

In terms of their topology structure, motif PML-STAT1-GCH1 forms a linear 021 motif with STAT1 being in the middle indicates that PML and GCH1 may be correlated, but any predictive effect from PML to GCH1 (or vice versa) is mediated by STAT1. In particular, the 021 motif topology indicates that STAT1 and GCH1 are independent of one another and are only associated with one another by their interactions with STAT1. Slightly different from the other two motifs is STAT1-LCK-NR1H3 motif that form three complete motifs and two linear motifs from the eight Basal-specific networks (Table 2.2). The inconsistency of the motif type and topology observed among node STAT1, LCK and NR1H3 could be caused the choice of threshold level. This work uses 1% thresholding which means any edge between STAT1, LCK and NR1H3 that are weaker than  $\pm 1\%$  will be excluded from the network, resulting in the change of the motif type (e.g. from type 030 to type 021). Nevertheless, the network motifs identified in this work reveal important NR associations between critical transcripts in Basal-like, providing a foundation for identifying oncogenic signalling pathways contributing to Basal-like breast cancer tumorigenesis.

#### 2.6.4 Limitations and future perspectives

While the results presented in this chapter revealed critical NR-based network modules in Basal-like breast cancer, NR transcriptional networks only reflect a fraction of the system-wide interactome, and further study to explore other functionally related subsets of genes (such as other transcriptional factors or kinases) could be pursued. Nuclear receptors were the focussed of this work. Widening the number of genes is expected to reveal other critical functional units underlying BLBC. The next challenge would be to understand the dynamics and communication between these critical network modules and potential identify critical cross-talks event that contribute directly to BLBC oncogenesis.

Another consideration is that the 3-node motifs demonstrated in this chapter only describe average associations (within a cluster of patients) and do not provide evidence of causations nor provide an explanation of the possible regulatory mechanisms. A correlation motif could be explained by different regulatory mechanisms (Figure 2.18). Further work should explore the possible regulatory mechanisms of the three critical motifs identified in this work. This could be achieved experimentally by carrying out cell-based assays or other biochemical assays, or kinetically by using ordinary differential equation models.



Figure 2.18: A correlation network motif and its possible regulatory mechanisms. A typical complete network motif identified from a correlation network is shown in the blue box. It could represent several different possible regulatory mechanisms, shown in the green box. The genes could be regulated in a feedback loop (left), or a feed-forward loop (middle), or by a common regulator (x), which is not part of the motif (right). Concept and figure were adapted from [220].

Transcription networks are examples of complex networks and revealing individual biomarkers or drug target candidates only provides limited insight and lacks a global perspective of the system-wide interactions [248, 249]. The work presented in this chapter not only revealed the critical genes (hubs) implicated in Basal-like subtype but also discussed the properties of these genes in the context of partial correlation networks. Compared to individual genes or biomarkers, the network motifs identified in this work represent a robust description of the transcriptional activities underlying Basal-like breast cancer and deserve further investigation [197]. This approach of integrating unsupervised clustering algorithms with network modelling is not restricted to breast cancer and could be tailored and reformulated to study other complex diseases [250–252].

# 2.7 Key findings

- The subtype stratification revealed by BHC showed a similar composition of tumours to the PAM50 subtype classification.
- Basal-like has distinct nuclear receptor expression patterns from Luminal subtypes; however, a fraction of Basal-like breast cancer was found to have Her2 characteristics, making the distinction between Basal-like and Her2 subtype ambiguous.

• The work presented in this chapter identified three hub (FOS, STAT1) associated 3-node motifs, representing critical network modules in Basal-like breast cancer.

# Chapter 3

# Gene Expression Clustering

### 3.1 Introduction

In this chapter, the second set of results produced by BHC will be considered. The second clustering strategy uses the transform of the dataset describes in Chapter 2, to cluster the NR genes according to the patient profiles. This approach groups NR genes into distinct gene clusters, with each cluster displaying a unique expression pattern. To make the analysis manageable, only results from the TCGA dataset (the largest RNA-seq-based dataset) was considered in this chapter. The same approach can be adapted and applied to METABRIC or other expression datasets. Although METABRIC used the older microarray technology (instead of RNA sequencing), an equivalent analysis from METABRIC would expect to reveal very similar results.

#### 3.1.1 Targeting Basal-like subtype

Basal-like breast cancer (BLBC) is characterised by an expression pattern similar to normal basal or myoepithelial cells of the breast that includes high expression of cytokeratin 5, 14, and 17 and proliferation-related genes [40]. BLBC is associated with a clinically aggressive nature and poor prognosis [40, 59, 253]. While molecular subtypes defined by gene expression profiles have provided prognostic and predictive values, the description of Basal-like subtype is not used routinely in the clinic [39]. This is because microarray-based transcriptional profiling requires fresh frozen tissue samples which is not always feasible in the clinic [65]. Instead, the clinicopathological classification uses immunohistochemical markers (ER, PR, Her2) to stratify patients into ER-positive, Her2-positive and triple negative subtypes. Triple negative breast cancer (TNBC) lacks the expression of hormone receptors (ER and PR) or Her2 receptor kinase. Bertucci *et al.* demonstrated that 71% of triple negative breast cancer (TNBC) were Basal-like, and 77% of Basal-like tumour had triple negative characteristics [254]. While the immunohistochemical definition of TNBC and transcriptional profiling-based definition of BLBC are not synonymous, due to the high overlap, the two terms (TNBC and Basal-like breast cancer) are sometimes, particularly in clinical reports, used interchangeably. Patients with tumours that represent TNBC characteristics do not benefit from more effective targeted therapies such as tamoxifen and aromatase inhibitors (ER) or trastuzumab (HER2 amplification) [255, 256]. Thus, individual or combinations of nonspecific cytostatic and cytotoxic approaches such as surgery, radiotherapy, chemotherapy are currently the only treatment options for tumours with TNBC characteristics [257].

In order to refine and improve BLBC/TNBC treatment options, much research has been focussing on revealing molecular characteristics associated with TNBC or BLBC subtype with the hope of identifying potential molecular targets. For example, VEGF pathway is over-activated in Basal-like, and several approaches have been proposed to inhibit this pathway [258]. This includes inhibiting the binding of VEGF to its cell surface receptors (e.g. bevacizumab, aflibercept) [259], directly blocking the receptor (e.g. ramucirumab) [260] or using small molecule multitarget tyrosine kinase inhibitors (e.g. sorafenib) [261]. While most of these therapeutic molecules have entered phase II clinical trials, their effect on TNBC/BLBC tumours is still under investigation [262]. Moreover, studies have also shown that EGFR is upregulated in most BLBC and TNBC tumours [70, 263, 264]. This observation encouraged and informed subsequent drug discovery research in developing EGFR inhibitors. Currently, there are two main types of EGFR inhibitors, including small molecule typosine kinase inhibitors (e.g. gefitinib and erlotinib) and monoclonal antibodies (e.g. cetuximab) [265]. However, clinical trials showed that targeting EGFR as a single target only results in modest benefit [93, 266–268]. While activation of other signalling pathways such as mTOR [269] and MEK [55] pathways have also shown to be aberrant in TNBC/BLBC tumours, whether targeting these pathways could improve clinical outcome in TNBC/BLBC patient needs to be investigated in further clinical trials. To date, BLBC tumours or tumours with TNBC characteristics remain a clinical challenge, and the molecular pathogenesis needs to be better understood. This chapter describes additional work where BHC was used to revealed expression patterns of nuclear receptors and their coregulators in different breast cancer subtypes, with an emphasis on Basal-like. From the gene

clustering results, 22 NR-associated genes were found to be highly expressed in BLBC, representing potential drug target candidates. These 22 genes were further evaluated and prioritised using a correlation network for subsequent analyses.

#### 3.1.2 Aims

The aims of this chapter were to:

- develop an approach for interpreting gene clusters revealed by BHC,
- identify the critical candidates that could be further explored at the protein level for drug discovery studies.

#### 3.1.3 Chapter overview

The steps for identifying critical candidate genes from BHC gene clusters are described in Figure 3.1 and the details of the methods and materials are described in Section 3.2. To examine the properties of gene clusters revealed by BHC, gene clusters containing known drug targets for non Basal-like subtypes were identified from corresponding analysis (detail described in Section 3.3.1). Gene clusters likely to contain potential drug targets for Basal-like patients were then identified. The members of these gene clusters were then compared to reveal the overlapping genes. Overlapping genes are referred to as basal candidate genes (BCGs) in this chapter. BCGs were then prioritised according to their topological property.



Figure 3.1: Overview of the chapter.

## 3.2 Methods

In order to identify gene expression pattern associated with Basal-like breast cancer, the second set of results produced by BHC are considered. The data and BHC clustering analysis are described in Section 2.2.

#### 3.2.1 BHC clustering and identification of Basal Candidate Genes

In order to reveal NR expression pattern that is unique to Basal, this chapter looked to analyse the second set of pairwise clustering results by BHC. In each comparison, the gene expressions were organised into clusters. Genes grouped into the same clusters are expected to exhibit similar expression patterns across observations (patients) and are likely to be involved in the same signalling pathway or be regulated by the same set of transcriptional factors [270, 271]. In BHC clustering, the z-score distributions of genes are discretised, and the peak of the distribution is called the marginal likelihood with two boundaries. Any value lower than the lower bound is represented in green. Any values greater than the upper bound is represented in red. The clustering only considers the signals (fraction of the data) between the two boundaries, which is repented in black. The area under the peak signal is estimated during the discretisation and is different for each pairwise analysis. Gene clusters that fall into the red upper bound (high expression) in Basal-like patients were extracted from each BHC comparison. The extracted gene lists were then compared, and genes found in all three comparisons were identified and referred to as "Basal candidate genes (BCGs)".

#### 3.2.2 Basal-specific correlation network

To further evaluate and prioritise the Basal candidate genes, a correlation network was constructed from the 98 Basal patients using the R/GeneNet library. Code used for network analysis were adapted from this GeneNet example (http://www.strimmerlab.org/software/ genenet/download/ecoli-net.R). The function R/GeneNet:network.test.edges was used to create a pairwise partial correlation matrix of all correlations, listed by order of magnitude. An empirical marginalised thresholding of 1% was used to reveal the most informative networks, after testing a series of different thresholding levels (35%, 25%, 15%, 10%, 5%, 2%, 1%, 0.5%). R/Graphiz was used to generate networks for visualisation. In the Basal specific network, each node in the network represents a gene and the lines between them represent a strong (top 1%) correlation. BCGs were identified and highlighted in the network. They were ranked according to their connectivity, defined as the number of connections that a node makes in a network. Hubs ,most connected nodes, were further explored.

# 3.3 Gene clustering results

While the patient clustering produced by BHC in the previous chapter provides useful insight into understanding the breast cancer heterogeneity, the resulting gene clusters from BHC are useful for identifying potential Basal drug targets for further drug discovery studies. Savage *et al.* have demonstrated that BHC is more likely to reveal biological relevant clusters (such as being responsible for a specific biological function or involved in a particular signalling pathway) compared to other classical clustering methods [184]. In this analysis, expression data of 171 nuclear receptor associated genes were obtained from four different breast cancer subtypes from TCGA (as described in Section 2.3). The 171 NR-associated genes were stratified into distinct clusters, with each cluster representing a distinct expression pattern. In the following BHC plot and dendrogram, each row represents a patient sample, and each column represents a gene expression. The expression level of each gene in each sample is relative to its marginal likelihood across the samples and is coloured according to the colour scheme shown under the colour plot. Green and red indicate expression levels below and above the marginal likelihood, respectively. The dendrograms on top of the colour plots show the hierarchical clustering of the gene clusters. In this section, the expression patterns associated with the critical gene clusters will be explored and described in details.

#### 3.3.1 Drug target assessment

The oestrogen receptor (coded by ER1), progesterone receptor (coded by PR) and Her2 (coded by erb-B2 Receptor Tyrosine Kinase 2 (ERBB2)) are three breast cancer biomarkers. Their status are used to determine a patient's prognosis and treatment [272]. Oestrogen receptors have been used as therapeutic targets, and several hormonal therapies (e.g. Tamoxifen) have been developed to target this receptor for treating Luminal type (ER-positive patients). While the effects of antiprogestogens in breast cancer are still under investigation, several studies have highlighted the potential clinical benefit of antiprogestin/progestin therapy in treating Luminal type [273, 274]. Anti-Her2 therapies such as trastuzumab have also shown great success in improving the survival of breast cancer patients with Her2-positive tumours [275]. Basal-like patients often lack these three biomarkers and therefore, do not respond well at all to these therapies.

In order to identify potential drug targets for Basal-like, and understand the signal pattern of existing drug targets, the gene group that has ER1, PR (drug target for Luminal A and Luminal B) and Her2 (drug target for Her2) were identified from the Basal-like vs Luminal A, Basal-like vs Luminal B and Basal-like vs Her2 comparisons, respectively. Figure 3.2 shows the gene clustering results produced by BHC. Ninety-eight Basal-like patients were compared to 231 Luminal A patients, 127 Luminal B patients and 58 Her2 patients, producing 18, 19, 11 distinct gene clusters respectively. These gene clusters present distinct expression signalling patterns and transcriptional behaviour.

Luminal drug targets, ER1 and PR, were both found in gene cluster 1 from both Basal-like vs Luminal A, Basal-like vs Luminal B analyses. Drug target for Her2-positive patients, ERBB2, was found in gene cluster 7 in Basal-like vs He2 comparison. Gene clusters that contain known drug targets are highlighted in blue in the corresponding dendrogram and colour plot in Figure 3.2. Importantly, all gene clusters that contain drug targets shown to be highly expressed (upperbounds, red) in the corresponding subtypes. In particular, cluster 1 from the Basal-like vs Luminal A analysis contains both ER1 and PR and is shown to be highly expressed in Liminal A. Likewise, the cluster that contains ER1 and PR from Basal-like vs Luminal B analysis is also highly expressed in Luminal B patients. Finally, ERBB2-containing gene cluster from Basal-like vs Her2, cluster 7, is also highly expressed in Her2. It is important to note that the BHC upper and lower bounds, the red and green, are different from gene expression heatmaps based on z-scores, and red in BHC colour plots represents highly expressed (upper bound) genes and does not represent overexpression.



Figure 3.2: Drug target containing gene clusters revealed by BHC. The gene clusters that contain drug target(s) for non-BLBC are highlighted in blue in the corresponding dendrograms and colour plots. Patient groups are also coloured and shown at the right of the colour plots, with Basal-like classes in orange, non-Basal classes in blue and groups that are difficult to defined (ambiguous group) in grey. Left: The 171 genes NR associated genes were stratified into 18 clusters in Basal vs Luminal A analysis. ER1 and PR are both in cluster 1, which is a gene cluster that is highly expressed in Luminal A. Middle: Genes were stratified into 19 clusters in Basal-like vs Luminal B analysis. ER1 and PR are both in cluster 1, which is a gene cluster that is highly expressed in Luminal B. Right: In Basal-like vs Her2 analysis, genes were stratified into 11 clusters. ERBB2 is in cluster 7, which is a gene cluster that is highly expressed in Her2.

#### 3.3.2 Clustering results

In Section 3.3.1, gene clusters containing drug target(s) revealed by BHC are highly expressed in the corresponding subtypes. From the gene clusters, it was presumed that the Basal-like drug target candidates are potentially those that fall in the higher bound (red) in Basal-like patients. Figure 3.3 shows the gene clusters that are highly expressed in Basal-like patients from each pairwise comparison. The gene clusters that are highly expressed in Basal are highlighted in orange in the dendrograms. In the Basal-like vs Luminal A comparison, the 171 NR-associated genes were stratified into 18 distinct clusters, of which cluster 2 and cluster 16 are highly expressed in Basal. In Basal-like vs Luminal B comparison, the 171 genes were stratified into 19 distinct clusters, and of which cluster 5, cluster 14 and cluster 15 are highly expressed in Basal-like patients. Finally, in Basal-like vs Her2 comparison, genes were stratified into 11 distinct gene clusters and of which cluster 1 and cluster 6 are highly expressed in Basal breast cancer. The highly expressed genes identified from each analysis are described in Appendix H.



Figure 3.3: Gene clusters in Basal-like subtype. Gene clusters that are highly expressed in Basal-like are highlighted in orange in the dendrograms and colour plots. Patient groupings are coloured at the right of the plots, with Basal-like classes in orange, non-Basal classes in blue and groups and those are difficult to defined (ambiguous group) in grey. Left: Genes stratified into 18 clusters in Basal vs Luminal A analysis, with cluster 2 and cluster 16 being highly expressed in Basal-like. Middle: Genes stratified into 19 clusters in Basal-like vs Luminal B analysis, with cluster 5, cluster 14, cluster 15, cluster 16 being highly expressed in Basal-like. Right: Genes stratified into 11 clusters in Basal-like vs Her2, with cluster 1 and cluster 6 being highly expressed in Basal-like.

# 3.4 Basal candidate gene identification

Genes in the clusters that are highly expressed in Basal-like were extracted and compared between the three BHC analyses. The Venn diagram in Figure 3.4 shows the three sets of genes extracted from each of the clustering, the intersections show the overlapping genes between the gene lists (the size of the circles is irrelevant to the size of gene clusters and do not represent the number of genes). A total number of 23, 32 and 44 genes were found to be highly expressed in Basal-like patients when comparing to Luminal A, Luminal B and Her2 patients, respectively. Twenty-two genes were overlapping between all three gene lists and are referred to as the Basal Candidate Genes (BCGs).



Figure 3.4: Venn diagram of the three sets of highly expressed genes extracted from the paired clustering analyses. Genes highly expressed in Basal from Basal vs Luminal A, Basal vs Luminal B and Basal vs Her2 analysis are shown in green, yellow and blue, respectively. The intersections are the overlapping gene lists, referred to as Basal candidate genes (BCGs).

## 3.5 Basal associated correlation networks

To prioritise the 22 BCGs, a Basal-specific correlation network was constructed from TCGA Basal patients to examine the connectivity of each BCGs. Each of the 22 BCGs was identified in a Basal correlation network (Figure 3.5) and are ranked according to their degree of connectivity (Table 3.1). The Basal correlation network was derived from all 98 Basal patients from TCGA using the same R packages and method described in Chapter 2.



Figure 3.5: Partial correlation network derived from Basal-like patients. Each node represents a gene, and each edge represents a partial correlation between genes. Edges are coloured with negative correlation in black and positive correlation in red. The numbers on the edges represent the strength of the correlations. Basal candidate genes are coloured in yellow, their corresponding degree are describe in Table 3.1

	BCGs	Connectivity	Biological function
1	STAT1	4	It is responsible for mediating cellular responses to inter-
			ferons, cytokines and other growth factors.
2	LCK	4	It can phosphorylate tyrosine residues of target proteins
			involved in the T-cell antigen receptor (TCR)-linked sig-
			nal transduction pathways.
3	NR3C2	3	It is a receptor for mineralocorticoidss (MC) and
			glucocorticoids (GC). Its activation leads to the expres-
			sion of proteins regulating ionic and water transports.
4	PPARGC1A	3	It acts as a transcriptional coactivator for steroid recep-
			tors and nuclear receptors, regulating genes involved in
			energy metabolism.
5	TMPRSS2	3	It is a serine protease that participates in proteolytic cas-
			cades in prostate. Its' over-activation has also been linked
			to prostate cancer metastasis.
6	RARB	3	It binds as heterodimers to their target response elements
			in response to their ligands, mediating cellular signalling
			in embryonic morphogenesis, cell growth and differentia-
			tion.
7	SIX3	2	As a transcriptional factor, it binds to target DNA se-
			quences, playing a crucial role in forebrain development
			and eye development.
8	NR2E1	2	It is a orphan receptor that binds DNA as a monomer
			to hormone response elements. It can regulate the ex-
			pression of another nuclear receptor, RAR. It is involved
			in the regulation of retinal development and essential for
			vision.
9	PPARA	2	It is a ligand-activated transcriptional factor, regulating
			lipid metabolism in the liver.

Table 3.1: Degree of each Basal candidate genes (BCGs) in the basal partial correlation network. Gene name abbreviations can be found in the bigining of the thesis.

80

10	TAF15	2	It codes for a RNA and ssDNA-binding protein that ini-
			tiate transcription and promoter recognition.
11	STAR	2	It is a transport protein that mediates the transfer of
			cholesterol within the mitochondria.
12	SOX9	2	it is a DNA-binding protein that regulate chondrocytes
			differentiation and skeletal development.
13	POMC	1	POMC is preproprotein that can be cleaved to give rise to
			different peptide hormones in response to different stim-
			ulation.
14	NR2C2	1	It is an important repressor for a number of nuclear re-
			ceptor signaling pathways, including thyroid hormone re-
			ceptor and estrogen receptor pathways.
15	PPARGC1B	1	It plays a is a transcriptional activator that can activate
			transcriptional activity of ER1, nuclear respiratory fac-
			tor 1 (NRF1) and nuclear receptor subfamily 3 group C $$
			member 1 (NR3C1) in the presence of glucocorticoids.
16	FABP6	1	It binds to fatty acids and bile acids, it is is involved in
			enterohepatic bile acid metabolism.
17	WT1	1	It is a DNA-binding transcriptional factor that are in-
			volved in regulating the cellular development and cell sur-
			vival.
18	BCL11A	1	It is a DNA-binding transcription factor that regulates
			gene expression via chromatin remodelling.
19	TCF3	0	It is a transcriptional regulator involved in the initiation
			of neuronal differentiation and mesenchymal to epithelial
			transition.
20	PAX6	0	It is a transcription factor involved in the development of
			the eye, nose, central nervous system and pancreas.
21	NR1I2	0	It is activated by variety of naturally occurring steroids,
			regulating genes involved in the metabolism and secretion
			of toxic compounds.

81

22	NPAS2	0	It is a transcriptional factor that is involved in the main-
			tenance of circadian rhythms.

While 22 BCGs were identified from the clustering analysis, transcription factor 3 (TCF3), paired box protein Pax-6 (PAX6), NR1I2 and neuronal PAS domain protein 2 (NPAS2) were not found in the Basal-specific correlation network. This suggests that, in TCGA, TCF3, PAX6, NR1I2 and NPAS2 are not strongly correlated to any other 170 NR associated genes. This could be because they are correlated to other sets of transcriptional factors or the correlations between TCF3, PAX6, NR1I2, NPAS2 and other genes are weak (below the 1% thresholds) in BLBC samples.

LCK and STAT1 are the most connected BCGs in the network with connectivity of four. Genes with high connectivity in a transcriptional correlation network assumed to be critical for regulating signalling pathways and cellular functions [276], and here controlling transcriptional events implicated in Basal-like subtype. Importantly, from the correlation network LCK is strongly correlated with STAT1. The positive correlation between LCK and STAT1 was further evaluated at the protein level, demonstrated in the next chapter.

# **3.6** Discussion and future perspectives

In this chapter, the second set of BHC results, where the NR genes were clustered according to the patient profiles, were analysed. To examine the gene clusters revealed by BHC and their associated expression signal patterns, known breast cancer drug targets for non-Basal-like subtypes (Luminal A, Luminal B and Her2) were identified from the corresponding analyses. Here, the drug targets (ER1 and PR) for Luminal type breast cancer were demonstrated in gene clusters that are highly expressed (upper bound) in Luminal type breast cancer. Similarity, the Her2 drug target (ERBB2) was found in the gene cluster highly expressed in Her2. To identify the Basal candidate genes, those genes in highly expressed clusters in Basal-like were identified and compared between the three analyses. Among these genes, twenty-two of them were overlapping in all comparisons and are referred to as Basal candidate genes (BCGs). The topology of the 22 BCGs was examined in a Basal only correlation network, with LCK and STAT1 having the maximal connectivity. The implication of this result from the gene clustering is discussed and considered. In cluster analysis the genes are grouped into clusters based on a defined similarity measure [176]. In BHC, the similarity measure represents similarity in expression patterns across samples (patients), and genes in the same BHC cluster are assumed to be functionally related or are regulated by a common set of transcriptional factors [184]. From the results, gene clusters that are highly expressed in Basal-like breast were identified. These gene clusters often have opposite expression behaviour in the Basal-like subtype compared to non-Basal subtype. For example, gene cluster 2 from the Basal-like vs Luminal A analysis are highly expressed in Basal-like and lowly expressed in Luminal A. This distinct expression pattern observed in Basal-like and non-Basal-like subtypes indicates that genes in these gene clusters might be involved in signalling pathways or processes over-activated in Basal-like.

Importantly, the two highly connected BCGs, LCK and STAT1, revealed from this chapter were also found in one of the three motifs identified in Chapter 2. The motif analysis from Chapter 2 suggests that LCK and STAT1 are highly connected (correlated) and enriched in Basal-like breast cancer networks. Whereas the gene clustering analysis presented in this chapter suggests that LCK and STAT1 are highly expressed in Basal-like breast cancer. It should be noted that, not all genes involved in the motifs were highly expressed in Basal-like breast cancer. This is indeed expected as the hub or hub-correlated genes described in Chapter 2 were defined by their connectivity within a given network, whereas the highly expressed genes are defined by their normalised expression magnitude. Hub genes that are not highly or differentially expressed in Basal-like breast cancer can be crucial for essential cellular functions and drugging such encoded proteins can potentially lead to unwanted toxicity. Whereas highly expressed genes that are not hubs might not have the same regulatory potential for regulating relevance signalling pathways as hub genes. The two clustering analyses suggest that LCK and STAT1 are both highly connected and highly expressed, highlighting the relevance of these two genes. LCK and STAT1 therefore present valuable targets that deserves further investigation.

While not the direction of this study, the results revealed from this analysis could be used for pathway enrichment analysis. Pathway enrichment comprises typically three main steps: identification of gene list of interest, determination of statistically enriched pathways, and visualisation and interpretation of the results [277]. The 22 BCGs identified in this chapter could be used as the gene list of interest, and the statistically enriched pathways analysis would identify biological pathways that are enriched in the 22 BCGs more than would be expected otherwise. Such analyses could provide insight into understanding the transcriptional activities underlying Basal-like breast cancer.

Another possible extension of this work would be to explore the prognostic value of BCGs and how they response to chemotherapies. The status of the 22 BCGs might provide a useful indication or prediction measure for patients' clinical outcomes. To investigate the potential of BCGs as prognostic markers, further survival analysis would need to be performed. For example, patients could be sub-dividing into BCG-high (above median) and BCG-low (below median) groups and their associating survival curves under difficult chemotherapy intervention could be plotted on a Kaplan-Meier plot. Cox proportional hazards regression can also be performed to describe survival. To ensure the representativeness of the survival analysis, large clinical datasets of patients biopsy and survival data would be required.

# 3.7 Key findings

- The BHC clustering results identified 22 genes (Basal candidate genes) that are highly expressed in Basal-like breast cancer.
- The 22 genes were prioritised using an network modelling approach, which suggested that STAT1 and LCK were the most critical BCGs given their high connectivity.

# Chapter 4

# Identification of Critical Protein-Protein Interactions

# 4.1 Introduction

As demonstrated in the previous chapter, LCK and STAT1 were the most connected Basal candidate genes (BCGs) and their mRNA expression correlated to each other in Basal-like breast cancer. In this chapter, the association between LCK and STAT1 at the protein level were explored using multiple online databases and bioinformatics tools.

#### 4.1.1 Protein-protein interactions

Proteins are large biomolecules composed of many amino acids linked via peptide bonds. Proteins are gene products that are essential for regulating a wide range of cellular processes, including cell growth, proliferation, signalling transduction and apoptosis. In order to facilitate these processes, proteins are required to communicate and bind to different interacting partners and metabolic. Protein-protein interactions (PPIs) are highly specific physical contacts between two or more proteins. They can be induced by different molecular interactions, including longrange electrostatic (polar) interactions, short-range dipole-dipole interactions and the effect of bulk waters driving the hydrophobic and aliphatic interactions.

While protein-protein interactions have many significant properties, such as modifying the kinetic properties of enzymes or allowing substrate channelling, this thesis will emphasise on the PPIs involved in signalling pathways. Signalling pathways transmit extracellular signals from
cell membrane receptors via signal transduction networks to the nucleus, requiring a series of sequential and transient associations and disassociations between upstream and downstream signalling proteins. Signalling PPIs are tightly regulated, and they are often subjected to variants in their sequence (isoform variations, mutations) affecting their surfaces or post-translational co-valent modifications (e.g. phosphorylation). Pawson and Nash were among the first to propose the idea that PPIs define the specificity in signal transduction in cells [278]. Their suggestion informed subsequent research into signalling PPIs in physiological and pathological states. It has been demonstrated that, in some cases, abnormal signalling of PPIs can lead to aberrant regulation of signal pathways and cause serious diseases, including breast cancer [279].

#### 4.1.2 Targeting protein-protein interactions

Modern drug discovery aims to identify new therapeutic agents that can selectively target disease-specific mechanisms or pathways [280]. The progress in molecular biology and genomic studies has improved our mechanistic understanding of human diseases and led to the identification of a large number of potential drug targets. Among these, protein-protein interactions represent an important class of molecular targets for therapeutic intervention, due to their central role of controlling cellular functions and processes [281].

While targeting PPIs for therapeutic intervention has historically been considered to be challenging due to their relatively large binding surfaces, substantial progress has been made in targeting disease-related PPIs using novel therapeutic agents, such as small molecules and peptide mimics [282, 283]. One of the most famous examples was the discovery of small-molecule inhibitors for targeting p53-MDM2 interaction as an anticancer strategy [284, 285]. Other oncogenic targets that have been successfully targeted and reached clinical development include BCl2 complexes, inhibitors of apoptosis proteins (IAP) and bromodomains [286]. These examples have shown that PPIs represent an attractive class of oncogenic targets, and modulating signalling PPIs is particularly relevant to the discovery of novel therapeutic agents for cancer treatment [287–289].

In this chapter, the mRNA correlations between LCK and STAT1 revealed from the previous chapter are explored at the protein level using several publicly available online databases and structural bioinformatics tools. Work presented in this chapter will uncover how LCK and STAT1 are associated with one another at the protein level and identify potential PPI targets for Basal-like breast cancer.

## 4.1.3 Aims

This chapter aims to achieve the following:

- uncovering the association between LCK and STAT1 at the protein level;
- identifying critical PPIs to target in Basal-like breast cancer;
- determing the structural characteristics of the proteins involved in the potential PPI targets.

#### 4.1.4 Chapter overview

The strategy presented in this chapter is shown in Figure 4.1 and the methods and databases used are summarised in Section 4.2. To uncover the association between LCK and STAT1 at the protein levels, two approaches were proposed (described in Section 4.3). Using online PPI databases, the first approach aimed to identify two intermediate protein binding partners between LCK and STAT1. However, most PPIs revealed from this first approach were only supported by a small number of experiments/pieces of literature, therefore not explored further. For this reason, a second approach was proposed to only consider high-confidence PPIs. The PPIs identified were represented in a PPI network and from this the PPIs were prioritised to identify those of interest. The structural characteristics of PPIs of interest were characterised using online databases and bioinformatics tools. The process and outcome of the structural examination are described in Section 4.4.



Figure 4.1: Chapter overview. The online databases or tools used in each step are shown in red.

In this chapter, all genes symbols are italicised, with all letters in uppercase (e.g. LCK) and all protein symbols are the same as the gene symbol but are <u>not</u> italicised (e.g. LCK).

## 4.2 Methods and databases

This chapter looked to explore the correlation between the two highly connected Basal candidate genes (BCG), LCK and STAT1. To reveal the association between LCK and STAT1 at the protein level, relevant information were gathered from multiple databases.

## 4.2.1 Protein-protein interaction networks

Protein-protein interaction networks were curated manually to demonstrate protein interactions around LCK and STAT1. The LCK and STAT1's binding partners were described in UniProtKB. To reduce the complexity of the network, we have only considered PPIs that have been demonstrated with more than 10 experimental evidence. Each line the in network represent a valid PPI and the thickness of the lines is weighted according to the numbers of experimental evidences.

#### 4.2.2 Protein structural information and SLiMs identification

Protein domains were identified and gathered from the Simple Modular Architecture Research Tool (SMART database). PPI binding sites were identified from the IntAct database (https:// www.ebi.ac.uk/intact/). All PPI binding sites were either curated from the literature or from data directly submitted to the IntAct. Peptide sequences from KIT and GAB1 were submitted to the eukaryotic linear motif (ELM) database (http://elm.eu.org) for short linear motifs (SLiMs) identification. All SLiMs identified have been experimentally validated by previous studies. The peptide stretches from KIT and GAB1 are in the intrinsically disordered regions of the proteins. Therefore, SLiMs that are expected to fall inside globular protein domains are excluded from the work.

## 4.3 LCK-STAT1 correlation

The previous chapter demonstrated that *LCK* and *STAT1* are the most connected Basal candidate genes (BCGs) with their expressions correlated to one another. To explore the association between *LCK* and *STAT1*, the gene (mRNA) expression of *LCK* and *STAT1* were assumed to be a proxy of protein expression and their association was assumed to be reflected at the protein level (either directly in a protein complex or indirectly through intermediate binding partners). While no direct interaction between LCK and STAT1 has been reported to the PPI databases (IntAct, Mint, UniprotKB), two approaches were developed to explore potential intermediate binding partners and PPIs surrounding LCK and STAT1.

#### 4.3.1 Approach 1: LCK-x-y-STAT1

Approach 1 considers two levels of intermediate binding partners between LCK and STAT1. Binary protein interactions were identified from three PPI databases, IntAct (https://www.ebi. ac.uk/intact/), Mint (https://mint.bio.uniroma2.it) and UniprotKB (https://www.uniprot. org). As LCK and STAT1 are strongly (top 1%) correlated to one another at the mRNA level; initially, two intermediate binding partners were expected to be identified to satisfy such PPI configurations:

#### • LCK-x-y-STAT1

where x is interacting with LCK and y; and y is interacting with x and STAT1. A total number

of 14 non-redundant pairs of x and y were found to satisfy the LCK-x-y-STAT1 PPI configuration from the PPI databases (Table 4.1).

Table 4.1: All the proteins that satisfy the LCK-x-y-STAT1 configuration. A total number of 14 combinations have been identified from the PPI databases. Full protein names can be found in the Abbreviation list at the beginning of the thesis.

	Protein 1	Protein 2 $(x)$	Protein 3 $(y)$	Protein 4
1	LCK	GAB1	EGFR	STAT1
2	LCK	HSP90AB1	EGFR	STAT1
3	LCK	LAT	EGFR	STAT1
4	LCK	MET	EGFR	STAT1
5	LCK	NR3C1	EGFR	STAT1
6	LCK	PTPN22	EGFR	STAT1
7	LCK	SYK	EGFR	STAT1
8	LCK	ZAP70	EGFR	STAT1
9	LCK	SH2D2A	ERBB2	STAT1
10	LCK	HSP90AB1	ERBB2	STAT1
11	LCK	PTPRC	ERBB2	STAT1
12	LCK	SYK	ERBB2	STAT1
13	LCK	KHDRBS1	STAT3	STAT1
14	LCK	SYK	STAT3	STAT1

Each LCK-x-y-STAT1 PPI configuration is comprised of three binary PPIs, LCK-x, x-y and y-STAT1. The detail of the binary PPIs that made up of the 14 PPI configurations is described in Table 4.2, including the types of interaction, and the number of experimental validations supporting a given PPI. Of the 14 LCK-x-y-STAT1 configurations, five types of binary interactions were identified, including direct interaction, physical contact, co-localisation, phosphorylation and dephosphorylation. Direct contact indicates that two proteins are in direct contact with each other. Physical contact is defined when two proteins are involved in the same protein complex and are in close proximity but not necessarily in direct contact with each other. Colocalisation is defined when a coincident occurrence of two proteins in a given subcellular fraction is observed with a low-resolution experimental method (e.g. image technique) from which a physical interaction between the two proteins can be inferred. Phosphorylation indicates a reversible covalent addition of phosphoryl group ( $PO_3^-$ ) on aspartic acid, cysteine, histidine, serine, threonine, tyrosine or arginine residues, while dephosphorylation indicates cleavage of a phosphoresidues from aspartic acid, cysteine, histidine, serine, threonine, tyrosine or arginine residues. While many potential intermediate binding proteins were identified, most of these PPI configurations are either poorly demonstrated (due to a lack of structural information or only confirmed by a small number of studies) or not applicable to Basal-like (e.g. PPI configurations that are mediating through ERBB2, as ERBB2 is absent in Basal-like). Moreover, some of the experimental validations (publications) have been conducted by the same group of researchers, and PPIs that are supported by a small number of validations are often only reported by a single research group. The lack of follow-up studies by other research group has suggested that the legitimacy of such PPIs might be controversial. For this reason, the PPIs identified from this approach were not pursued further, but considered and acknowledged as reference PPIs.

Table 4.2: Details of the 14 LCK-x-y-STAT1 PPI configurations. Each LCK-x-y-STAT1
configuration is composed of three binary PPIs: LCK-x, x-y and y-STAT1. The interaction type
associated with each binary PPI is described. The number of experimental evidence supporting
each binary PPI are shown in brackets.

PPI	LCK-x	x-y	y-STAT1		
EGFR-mediating PPIs					
LCK-GAB1-EGFR-STAT1	Direct interaction	Physical association	Physical association		
	(10)	(3)	(3)		
LCK-HSP90AB1-EGFR-STAT1	Physical association	Physical association	Physical association		
	(2)	(8)	(6)		
LCK-LAT-EGFR-STAT1	Phosphorylation	Physical association	Physical association		
	(2)	(2)	(6)		
LCK-MET-EGFR-STAT1	Direct interaction	Physical association	Physical association		
	(3)	(8)	(6)		
LCK-NR3C1-EGFR-STAT1	Physical association,	Physical association	Physical association		
	Co-localisation	(2)	(6)		
	(6)				
LCK-PTPN22-EGFR-STAT1	Dephosphorylating,	Physical association	Physical association		
	Direct interaction	(2)	(6)		
	(6)				
LCK-SYK-EGFR- STAT1	Physical association	Direct interaction	Physical association		
	(7)	(6)	(6)		
LCK-ZAP70-EGFR-STAT1	Phosphorylation,	Direct interaction	Physical association		
	Physical association	(2)	(6)		
(2)					
ERBB2-mediating PPI configurations(Not applicable to Basal)					
LCK-SH2D2A-ERBB2-STAT1	Physical association	Direct interaction	Direct interaction		
	(12)	(2)	(3)		
LCK-HSP90AB1-ERBB2-STAT1	Physical association	Physical association	Direct interaction		
	(2)	(3)	(3)		
LCK-PTPRC-ERBB2-STAT1	Dephosphorylation	Dephosphorylation	Direct interaction		
	(7)	(2)	(3)		
LCK-SYK-ERBB2-STAT1	Physical association	Direct interaction	Direct interaction		
	(7)	(7)	(3)		
STAT3-mediating PPI configurations					
LCK-KHDRBS1–STAT3- STAT1	Physical association	Physical association	Physical association		
	(5)	(8)	(3)		
LCK-SYK-STAT3-STAT1	Physical association	Physical association	Physical association		
	(7)	(2)	(3)		

#### 4.3.2 Approach 2: LCK and STAT1 PPI network

To pursue the study further, Approach 2 aimed to identify high-confidence PPIs that have been demonstrated by more than 10 experiments. Importantly, the second approach does not make an assumption regarding the number of intermediate binding partners between LCK and STAT1 (in Approach 1 the constraint was two partners only). Binary interactions that are associated with LCK and STAT1 were identified from the same PPI databases as Approach 1, but only interactions that have been confirmed by more than 10 experimental studies were considered. The resulting PPI networks are shown in Figure 4.2, with each node representing a protein entity, and the edges between the nodes representing high-confidence PPIs. The thickness of the edges is weighted according to the number of supporting experimental validations for each given PPI. The LCK associated PPI network is shown in blue on the left of Figure 4.2, the STAT1 associated PPI network is shown in green on the right of Figure 4.2. Hub proteins from each network are coloured in red.



Figure 4.2: Illustration of LCK (blue) and STAT1 (green) PPI networks. The thickness of the edges is weighted according to the number of experimental evidence of each valid PPI. The numbers next to the lines indicate the minimum and the maximum number of pieces of evidence of the PPI networks (min=10, max=43). The hubs of each network are coloured in red. ERBB2 (Her2) protein is absent in BLBC and potentially lost PPIs around ERBB2 (Her2) are shown in dotted lines.

In the LCK PPI network (Figure 4.2 blue), GAB1 and KIT are the hub proteins, representing critical positions in the network. Interestingly, GAB1 and KIT share multiple binding partners (e.g. PIK3R1, PIK3R2, PIK3R3, RAS P21 protein activator 1 (RASA1), protein tyrosine phosphatase non-receptor type 11 (PTPN11), phospholipase C Gamma 1 (PLCG1)). On the

other hand, the STAT1 PPI network (Figure 4.2 green) represents a centralised structure with EGFR being the hub. While overexpression of EGFR is frequently seen in BLBC [290], clinical trials of EGFR inhibitors for BLBC have failed due to low response rates [291].

Figure 4.2 also revealed the roles of ERBB2(Her2) and ERBB3(Her3) in bridging the LCK and STAT1 networks, providing an elucidation of the strong LCK-STAT1 correlation observed in the previous chapters. However, with the majority of Basal-like patients lacking ERBB2 protein expression, it is assumed that PPIs around ERBB2 (shown in dotted lines in Figure 4.2) are potentially lost or disrupted in BLBC, and therefore are not suitable drug target candidates.

Although studies have suggested that ERBB3 plays a vital role in regulating downstream oncogenic signalling pathways [292–294] and its activation is thought to be associated with drug resistance causing treatment failure in cancer therapy [295–298], results from Luhtala *et al.* showed that aggressive breast cancer subtypes such as Basal-like are more likely to have low expression of ERBB3. [299].

As previous studies suggest that EGFR, ERBB2 and ERBB3 are either lowly expressed in Basallike or not a suitable drug target for BLBC, the PPIs around these proteins were eliminated and not considered further. For this reason, a decision was made to focus on the interactions around hub proteins (KIT and GAB1) from the LCK network only. This decision is important as it ensures the attention is on or around the hub proteins.

## 4.4 Putative binding regions

In this section, the importance of KIT and GAB1 in breast cancer will be discussed, followed by several structural bioinformatics analyses to reveal and explore the putative binding sites of the KIT and GAB1 associated PPIs.

#### 4.4.1 KIT in breast cancer

KIT, also termed c-KIT or CD117, is a transmembrane receptor tyrosine kinase that is involved in multiple intracellular signalling pathways, regulating several important physiological processes (Figure 4.3). KIT can be activated by binding to its' ligand, stem cell factor (SCF) or also known as mast cell growth factor. Once activated, KIT undergoes dimerization and auto-phosphorylation. These phosphorylated sites in KIT then recruit downstream proteins to the cell membrane, regulating signalling pathways including PI3K/AKT and Ras GTPase (Ras)/MEK/MAPK pathways. The PI3K/AKT signalling pathway is known to play a role in carcinogenesis, promoting tumour survival, and growth [300, 301]. It is also amongst the most frequently dysregulated signalling pathway in Basal-like breast cancer [302].

In normal physiological circumstances, KIT is mainly expressed in hematopoietic cells [303]. It has been demonstrated that KIT plays an important role in hematopoiesis [304], fertility [305, 306], and melangenesis [307]. Deregulation of KIT (e.g. gain/loss of function, overexpression, and mutations) has been seen and linked to several cancers including breast cancer [71, 308–310]. While the significance of KIT in breast cancer in general, remain controversial [71, 311, 312], several studies have shown that KIT is more likely to be overexpressed in Basal-like breast cancer [313–315]. Previous research has also shown that KIT-positive Basal-like has a significantly worse prognosis than KIT-negative Basal-like [71, 315, 316]. This observation alone highlights the therapeutic potential of KIT-related interactions for a treatment of Basal-like breast cancer.



Figure 4.3: **KIT signalling pathways.** KIT, is activated upon binding to stem cell factor (SCF). Activation of KIT initiates multiple downstream signalling pathways, including the MAPK (mitogen-activated protein kinase, PI3K/AKT(Protein kinase B,), PLC $\gamma$  (phospholipase C- $\gamma$ ) and JAK/STAT (Janus kinase–signal transducers and activators of transcription) pathways. Activation of these pathways has been associated with cellular proliferation, differentiation and survival. Figure and content recreated from [317]. textbfSRC: proto-oncogene tyrosine-protein kinase; Lyn: tyrosine-protein kinase Lyn; DAG: diacylglycerol; IP3: inositol trisphosphate; BAD: BCL-2-associated death promoter; mTOR: mammalian target of rapamycin; GRB2: growth factor receptor-bound protein 2; SOS: son of sevenless; RAF: rapidly accelerated fibrosarcoma; ERK: extracellular-signal-regulated kinase; P: phosphorylation.

#### 4.4.2 GAB1 in breast cancer

Grb2-associated binding protein 1 (GAB1) is an adapter protein that can transduce signals from a variety of activated tyrosine kinases, including tyrosine-protein kinase MET (MET) and the epidermal growth factor receptor (EGFR). As a member of the IRS1 (Insulin receptor substrate 1)-like multi-substrate docking protein family, GAB1 play an important role in mediating signalling pathways that have been known to be involved in tumorigenesis, such as PI3K/Akt(Protein kinase B) and MET/hepatocyte growth factor (HGF) signalling pathway (Figure 4.4). GAB1 has a pleckstrin homology (PH) domain at its N-terminal, which can bind to phosphatidylinositol 3,4,5 trisphosphate (PIP3) and localise GAB1 to the cell plasma membrane, in the vicinity of activated tyrosine kinases [318]. GAB1 can be recruited to activate tyrosine kinases via direct interactions or indirect interactions involving growth factor receptor-bound protein 2 (Grab2) as a bridging protein [319]. GAB1 contains multiple phosphorylated tyrosine sites that can be served as binding sites for src homology 2 domain-containing protein effectors, such as the protein tyrosine phosphatase (Shp2), the phosphatidylinositol 3-kinase (PI3K) p85 regulatory subunits and the adaptor proteins Crk and Nck [320, 321].

GAB1 has been shown to be involved in tumour proliferation and metastasis in multiple tumours, including head and neck squamous cell carcinoma and colorectal cancer [322–327]. In breast cancer, a recent study done by Wang et al. showed that GAB1 is overexpressed in primary breast cancer clinical samples and is significantly upregulated in metastatic Basal-like breast cancer, particularly in HER2 and TNBC subtypes [328]. Their findings suggest that GAB1 might be a potential biomarker for breast cancer metastasis and inhibiting GAB1 or associated pathways might reduce the risk of breast cancer metastasis. In this work, the protein interactions associated with GAB1 will be explored and characterised.



Figure 4.4: **GAB1 signalling pathways.** GAB1 binds to activated tyrosine kinases (e.g. MET and EGFR) via direct interactions or indirect interactions involving GRAB2 as a bridging protein. GAB1 mediates multiple signal transduction pathways, including PI3K/AKT, PI3K/FAK (focal adhesion kinase) and RAS pathway. Figure and content recreated from [329]. **EGF**: epidermal growth factor; **HGF**: hepatocyte growth factor.

#### 4.4.3 Putative binding sites

As previously demonstrated, KIT and GAB1 share multiple binding partners, including PI3K regulatory subunits (PIK3R1, PIK3R2, PIK3R3), RASA1, PTPN11 and PLCG1. To explore the coordinated interactions between these proteins, the putative binding details of these PPIs were identified and summarised from the previous studies (Table 4.3). Considering the results from the previous work, we can see that while the binding sites for some of the PPIs remain unclear, a general binding pattern can be observed. Specifically, short stretches of sequences from KIT or GAB1 were shown to interact with the Src Homology 2 (SH2) domains from their shared binding partners.

Table 4.3: Summary of previously reported binding sites for KIT and GAB1 PPIs.PLA: proximity ligation assay; Co-IP:Coimmunoprecipitation; PPhase: phosphatase assay.

Idd	KIT/GAB1 binding regions	Partner binding regions	Detection method	References
	KIT protein-pro	tein interactions		
KIT-PIK3R1	AA714-727	SH2 domain	Co-IP	[330 - 332]
			PLA	
KIT-PIK3R2	AA714-727	SH2 domain	Co-IP	[331]
KIT-PIK3R3	Unknown	Unknown		I
KIT-RASR1	Unknown	Unknown		ı
KIT-PTPN11	Unknown	Unknown	ı	ı
KIT-PLCG1	AA929-942	SH2 domain	Co-IP	[331]
	GAB1 protein-pr	otein interactions		
GAB1-PIK3R1	AA440-453	SH2 domain	Co-IP	[333]
	AA465-478			1
	AA582-595			
GAB1-PIK3R2	Unknown	Unknown	·	ı
GAB1-PIK3R3	Unknown	Unknown	ı	ı
GAB1-RASR1	AA310-323	SH2 domain	Co-IP	[334]
GAB1-PTPN11	AA620-633	SH2 domain	Co-IP	[335]
	AA651-665		PPhase	
GAB1-PLCG1	Unknown	Unknown	ı	I

## 4.4.4 Prediction of structural domains in LCK and GAB1's binding partners

To further investigate the PPIs of interest, the predicted structural domains in PIK3R1, PIK3R2, PIK3R3, RASA1, PTPN11, PLCG1 were identified from SMART (Simple Modular Architecture Research Tool) database [336]. Interestingly, all six KIT and GAB1's shared binding proteins were predicted to have two SH2 domains each, with one being closer to the C-terminal (C-SH2) and one being closer to the N-terminal (N-SH21) (Figure 4.5). Their corresponding positions in each protein are described in Table 4.4.



Figure 4.5: Schematic representation of KIT and GAB1's binding partners. Signalling domains are illustrated using the colour, symbol and domain abbreviations from the SMART database. Each protein contains an N-SH2 and a C-SH2 domains highlighted in red boxes. SH2: Src Homology 2 domain; SH3: Src Homology 3 domain; RhoGAP: Rho GTPase activating protein domain; PH: Pleckstrin homology domain; C2: Calcium-binding domain; PTPc: Protein tyrosine phosphatase, catalytic domain; PLCXc: Phospholipase C, catalytic domain (part); domain X; PLCYc: Phospholipase C, catalytic domain (part); domain Y.

Protein (UniprotID)	N-SH2 location	C-SH2 location
PIK3R1 (P27986)	333-428	624-718
PIK3R2 (O00459)	330-425	622-716
PIK3R3 (Q92569)	65-160	358-452
RASR1 (P20936)	173-282	283-341
PTPN11 (Q06124)	1-103	104-217
PLCG1 (P19174)	528-661	528-661

Table 4.4: Location of the N-SH2 and C-SH2 domains in KIT and GAB1's binding partners.

## 4.4.5 SH2 domain alignment

To compare and explore the similarity of the N-SH2 and C-SH2 domains from LCK and GAB1's shared binding proteins, their sequences were aligned using a multiple sequence alignment tool 'Align' from Uniprot (https://www.uniprot.org/align/), which used the Clustal Omega algorithm to align sequences [337, 338]. The alignment results were downloaded in FASTA format, which was then visualised using the NCBI Multiple Sequence Alignment Viewer (MSA, v1.14.0) (https://www.ncbi.nlm.nih.gov/projects/msaviewer/). Amino acids were coloured according to a property-based colouring scheme (RasMol scheme), with acidic residues coloured in red, basic residues coloured in blue, non-polar residues are in yellow and polar, uncharged residues are in green. The alignment results (Figure 4.6) suggest that, among the six SH2-containing proteins, N-SH2 domains show a higher level of conservation than the C-SH2 domains, with 12 identical positions identified in N-SH2 and only two were identified in C-SH2. The alignment results also show that both N-SH2 and C-SH2 domains are more conserved among PI3K regulatory subunits (PIK3R1, PIK3R2 and PIK3R3) than in the other three proteins. This was not a surprising outcome, as PIK3R1, PIK3R2, and PIK3R3 are three variants of the PI3K regulatory subunits and are expected to be isoforms, functionally and structurally related.



N-SH2 domains

Figure 4.6: Sequence alignment of N-SH2 and C-SH2 domain from PIK3R1, PIK3R2, PIK3R3, RASR1, PTPN11, PLCG1. An asterisk (\*) indicates positions that are 100% conserved in the alignment; a colon (:) indicates conservation between groups of strongly similar properties (same colour code but not necessarily the same amino acids); fullstop (.) indicating weaker conservation between groups of weakly similar properties (5/6 amino acids have the same colouring code). RasMol colour scheme was used to colour the amino acids according to their properties. Acidic amino acids are in red; basic amino acids are in blue; non-polar amino acids are in yellow and polar, uncharged amino acids are in green.

## 4.4.6 Structural prediction of KIT and GAB1

Predicted structural domains were also identified for KIT and GAB1 (Figure 4.7) from SMART and Gene3D database. KIT comprised of five extracellular immunoglobulin domains and two intracellular catalytic domains (phosphorylase kinase and phosphotransferase domains). GAB1 is predominantly disordered and only comprised one structured domain. The prediction shown in Figure 4.7 indicates that most of the SH2-interacting short sequences in KIT and GAB1 are located in the intrinsically disordered regions (IDRs) of the proteins. IDRs are flexible regions of a protein that lacks a stable tertiary structure; their flexibility often allows them to adapt to various binding conformational requirements. The functionality of IDRs is attributable to short linear motifs (SLiMs) [339], equivalent to expose linear epitopes recognises by monoclonal antibodies. SLiMs are short fragments of sequences (typically 3-10 amino acids in length) that can be recognised by various classes of protein domains, giving rise to a wide spectrum of functionality. [340–342].



Figure 4.7: Predicted structural domains in KIT and GAB1. Structural prediction was obtained from Gene3D (http://gene3d.biochem.ucl.ac.uk/Gene3D/).

#### 4.4.7 SLiMs implicated in IDRs

To explore the functionality of the short stretches of sequences from KIT and GAB1, the sequences were scanned to match the defined SLiMs from the Eukaryotic Linear Motif (ELM) (http://elm.eu.org). Five SLiMs were found from KIT residues 714 to 727 (Fig 4.8A), including an N-terminal degron, two phosphorylation sites, one sumoylation site and tyrosine signal sorting motif. Seven SLiMs (one IAP binding motifs, one forkhead-associated domain (FHA) binding motif, three SH2 binding motifs, one phosphorylation site and a tyrosine signal sorting motif) were found from KIT residues 929 to 942 (Fig 4.8B). Six SLiMs (one phosphodegron motif, one phospho-dependent Cks binding motif, one phospho-dependent FHA binding motif, one Atg8 binding motif and two SH2 binding motifs) were found in GAB1 residues 310 to 323 (Fig 4.8C). Four motifs, including an IAP binding motif, a SH2 binding motif, Src homology 3 (SH3) binding motif and a tyrosine signal sorting motif, were found in GAB1 residues 440-453 (Fig 4.8D). Six SLiMs, including an N-terminal degron, a WW binding motif, a SH2 binding motif, a SH3 binding motif, TRAF2 binding motif and a tyrosine signal sorting motif, were found in GAB1 residues 465-478 (Fig 4.8E). Five SLiMs (one IAP binding motif, one SH2 binding motif, one SH3 binding motif, one CK1 phosphorylation site and one tyrosine signal sorting motif) were found in GAB1 residues 582 to 595 (Fig 4.8F). Eight SLiMs, including one N-terminal degron, two MAPL docking motifs, two Atg8 binding motifs, one SH2 binding domain, one immunoreceptor tyrosine-based motif and a tyrosine-signal sorting motif were found from GAB1 residues 620 to 633 (Fig 4.8G). Six SLiMs, including two types of Atg8 binding motifs, two types of SH2 binding motifs, one immunoreceptor typosine-based motif and a typosine

signal sorting motif were found in GAB1 residues 652 to 655 (Fig 4.8H).

The results from the SLiM examination showed that the short stretches of sequences in KIT and GAB1 consist of multiple SLiMs with most having one or more SH2 domain binding motif. This suggests that these sequences are likely to be recognised by the SH2 domains in their binding partners. Consequently, in agreement with results described in Section 4.4.3, the KIT and GAB1 PPIs are expected to be mediated through SH2 domains.



Figure 4.8: The predicted SLiMs from the short peptide sequences in KIT and GAB1. Peptide sequence and corresponding predicted SLiMs from (A) KIT residues 714 to 727, (B) KIT residues 929 to 942, (C) GAB1 residues 310 to 323, (D) GAB1 residues 440 to 453, (E) GAB1 residues 465 to 478, (F) GAB1 residues 582 to 595, (G) GAB1 residues 620 to 633 and (H) GAB1 residues 652 to 665. Top two rows in the tables show the sequence number and the corresponding amino acids of the peptides, respectively. All motifs listed are non-redundant. CK: Creatine kinase. IAP: Inhibitors of apoptosis proteins. ATG8: Autophagy-related protein 8. SH2: Src Homology 2. SH3: Src Homology 3. FHA: forkhead-associated domain. MAPK: mitogen-activated protein kinase.

## 4.5 Discussion

This chapter describes the process of identifying and characterising critical protein-protein interactions using several bioinformatics platforms. This includes revealing knowledge-based protein-protein interaction networks around LCK and STAT1 and highlighting the molecular interaction potential of the critical PPIs in the networks. In this chapter, 12 critical PPIs, between KIT, GAB1 and their shared binding partners, were revealed from the PPI network. Their putative binding regions and structural characteristics were highlighted. Analyses presented in this chapter suggest that these 12 PPIs are likely to be mediated by SH2-peptide binding. Here, criticisms and possible extensions of this work will be discussed in Section 4.5.1 to Section 4.5.3, and followed by a thought experiment, where different binding models of the PPIs were explored (Section 4.5.4).

#### 4.5.1 Gene-protein correlation

In this work, the gene (mRNA) expression of LCK and STAT1 were assumed to be the first proxy of protein expression and their strong association was assumed to be reflected at the protein level. While many biomedical studies make the same assumption that transcriptional expression is correlated to the protein level, some have argued that the gene-protein correlation is not always strong and can vary hugely depending on the system examined [343, 344]. It should be appreciated that transcription and translation are complex processes and indeed are affected by multiple factors [345]. The ongoing discussion highlights at least three factors that could influence the strength of a gene-protein correlation. Firstly, a wide range of posttranslational mechanisms that are involved in turning mRNA into proteins has not been well understood. This poses a significant challenge in predicting protein concentration from mRNA expression. Secondly, mRNA and protein stability are also a significant factor that could hugely affect the gene-protein correlation observed. Vogel et al. stated that degradation of mammalian mRNAs is much faster than mammalian proteins, with an average half-life of 2.6-7 hours versus 46 hours respectively [345–347]. Moreover, their stabilities are also affected by several other factors (e.g. environmental factors, molecular modifications etc.), that could lead to a weak correlation between mRNA and protein levels. Finally, both mRNA and protein expression experiments contain some levels of natural and manufactured systematic noise which could result in bias analysis [348, 349]. There is a continued effort to both describe and eliminate reduce

expression noise [350]. Nevertheless, although factors contributing to the mRNA-protein translation process are still being understood, mRNA expression data is still useful for identifying potential candidates for further investigation at the protein level. In particular, gene expression provides an indication of whether a protein is present in a system and approximately what level of protein expression is to expected.

In human, the LCK gene is differentially transcribed from two promoters, the proximal promoter and the distal promoter, separated by approximately 15 kb [351]. A number of previous studies showed that in T cells, activation of T lymphocytes can lead to down-modulation of LCK mRNA expression [352, 353]). This change in LCK mRNA expression was thought be to be caused by transcriptional and post-transcriptional regulations, especially a decrease of LCK gene transcription and mRNA stability during the early phase of the transcriptional process. Paillar and Vaquero also showed that the regulation of LCK is affected by a protein synthesis inhibitor cycloheximide [352]. While their results suggested cycloheximide is able to inhibit transcriptional activators that might control the expression of mRNA, they found that cycloheximide can massively increase the stability of LCK mRNA. The hypothesised that the stabilising effect induced by cycloheximide is so strong that it outweighs the inhibitory effect on the transcriptional activators.

STAT1 protein is a transcriptional factor that regulates expression of many target genes. It is evidenced that phospho-STAT1 protein can regulate the transcription of its own gene, representing a positive autoregulation [354]. While the complete regulation mechanism of human STAT1 remains partially understood, a recent study identified a novel distal regulatory element that mediates the positive feedback regulation of the STAT1 gene expression [355]. It is hypothesised that the positive feedback regulation of the STAT1 gene is interferon(IFN)-mediated and can be regulated via at least three pathways [355]. The hypothesis is that type I and II IFNs can bind and activate their corresponding receptors and this activation would then trigger phosphorylation of Janus kinases (JAK)and STATs [356, 357]. Phosphorylated STAT1 can then facilitate the formation of downstream protein complexes i.e. IFN-gamma-activated factor, STAT1–STAT2–IRF9(interferon regulatory factor 9) heterotrimer) [358, 359]. These complexes then translocate to the nucleus and bind to the STAT1's regulatory promoters and regulate STAT1 mRNA expression [360, 361].

A recent study done by Tang et al. showed that while in breast tumours, the tumour transcrip-

tome does not fully reflect the tumour proteome, the mRNA-protein abundance correlation is significantly higher in breast tumour tissues than the adjacent non-cancerous tissue [362]. They also showed that the increased concordance between mRNA and protein expression was associated with aggressive subtypes, including Basal-like, and decreased patient survival. High mRNA-protein correlation was observed particularly in proteins involved in disease metabolic pathways, suggesting enhanced protein-mRNA coupling in cancer metabolism.

These studies demonstrated that gene regulation is a rather complex process and is tightly regulated by multiple transcriptional factors. Different models of regulation mechanism, the stability of mRNA and the system examined, can all impact the mRNA-protein correlation. Instead of using mRNA expression as a first proxy of protein expression, mRNA-protein correlation of interest can also be examined from online databases or other large cancer-omics projects, such as The Human Protein Atlas (https://www.proteinatlas.org), Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle), Dependency Map (https://depmap.org/portal/). However, this should be examined with cautious, as tumour heterogeneity within a dataset and the variations in technology used can all have an effect on mRNA-protein concordance.

#### 4.5.2 Reliability of protein-protein interactions and target validation

Another criticism of this work is how the binary PPIs were selected for inclusion in the PPI networks. A constraint of at least 10 experimental pieces of evidence were used in Approach 2 to eliminate false-positive PPIs and to ensure reliable and a meaningful correlation for further investigation. However, the criteria used in this work is solely based on the number of available pieces of experimental evidence and did not consider other factors. To include the highest possible quality PPIs, future work should consider the experimental detection method used, and the type of interaction associated with a given PPI. Different detection methods are associated with different levels of specificity, for example biochemical (e.g. Co-IP) and biophysical (e.g. mass spectroscopy) methods generally provide more specific detection than image techniques (e.g. super-resolution microscopy). Similarly, types of association can also provide a good indication of the reliability of a given PPI. Several protein interaction confidence assignment schemes have been proposed, with each having their own unique parameters and weighting design. While Suthram *et al.* described a comprehensive comparison between different confidence scoring schemes [363], more recent confidence assessment methods had been developed. Future work could explore different options and choose an appropriate confidence assignment scheme for the data (e.g. some confidence schemes are designed for specific conditions: species, disease, cell/tissue type) to ensure the quality of the analysis.

Another limitation of this work is that the PPI information is only collected from a number of chosen databases (interaction information from IntAct, Mint and Uniprot; structural information from SMART and Gene3D). While the databases considered in this work are wellestablished and publicly available, it is likely that the information summarised in this work does not represent "the entire scientific literature". Future work can consider exploring metadatabase platforms, such as Pathguide, to integrate wider contents. Pathguide is a useful meta-database platform that provides an overview of more than 700 online biological pathway and network databases, including 320 PPI databases [364]. Systematically explore the PPIs of interest using the databases on Pathguide in the future is likely to reveal additional information that isn't available on the more commonly used databases. However, content aggregation from Pathguide should be proceed with cautious, as the databases listed on Pathguide are maintained by different scientific groups from different countries, and the definitions and quality control might vary from database to database.

Exploring the prognosis values of the potential drug targets identified in this chapter would be a valuable analysis. However, given the data availability it wasn't feasible to perform a statistically meaningful (p-value < 0.05) analysis at the time of research (2020). This was mainly due to two reasons: firstly, the number of Basal-like samples with protein expression data is low and protein expression data isn't available for all proteins; secondly, not all samples had associated followed up clinical metadata. With the recent technical advances in mass spectrometry, it is expected that the proteomic capacity will increase, and larger proteomic dataset will become available [365]. Evaluating large proteomic studies in the future will enable further validation on the prognosis values of the proteins of interest. Alternatively, exploring the prognosis values of relevance mRNA could also provide a level of indication. However, such approach would need to consider how well the mRNA-protein translation rate is for the target of interest, and therefore, how representative the mRNA expression is for the encoded protein, as discussed in Section 4.5.1. Several online bioinformatic tools have been developed for the purpose of easier and more accessible prognostic validation. For example, Kaplan-Meier Plotter, integrated survival data and expression data for more than 20 types of cancers from a wide range of open-access databases (original development described in [366], most recent update described [367]). Using such online survival analysis tool would allow rapidly assess of the prognostic value of the target of interest.

#### 4.5.3 Proteins with tandem SH2 domains

From the structural prediction analysis shown in Figure 4.5, all six KIT and GAB1's shared binding partners have two SH2 domains, arranged in tandem. Pawson suggested that SH2containing proteins often contain other signal-transducing domains, such as multiple SH2 domains, and SH3, PH domains [368]. The various combination of these domains is thought to increase a protein's binding specificity, as the respective order and spatial distribution of the domains would require a very specific topology for a high-affinity ligand or binding partner to interact [369]. This was in agreement with the observation reported by Ottinger *et al.*, in which they suggested that proteins with tandem SH2 domains show more significant binding specificities than those with single SH2 domain [370]. However, all these studies were done in the late 90s to early 2000s when SH2 domain was first discovered, and currently, there is a lack of more sophisticated discussion on the roles of N-SH2 and C-SH2 domains in different signalling proteins.

In this chapter, the sequence similarities of the tandem SH2 domains in the six shared binding proteins were explored. The conservation examination reveals a higher degree of conservation in N-SH2 than in C-SH2 domains, suggesting that while N-SH2 might be responsible for recognising more general feature (shared within a family protein), C-SH2 domain may have evolved to recognise more specific motifs and provide specificities of molecular recognition. While it is not the focus of this thesis, the results presented in this chapter have highlighted the importance of signalling proteins with tandem SH2 domains and exploring the functions and evolutionary relationship between N-SH2 and C-SH2 domains in different signalling proteins could be an interesting avenue for future research.

#### 4.5.4 Tandem SH2 protein binding models

The results presented in this chapter demonstrated that the short stretches of sequences in KIT and GAB1 are likely to bind to the SH2 domains in their binding partners. With SH2 domains being phosphorylated tyrosine (pY) binding domain, it was assumed that my chosen PPIs of interest are all phosphorylated dependent. While the structural characteristics of the proteins did not provide information about the binding dynamic or mechanism relationship of the interactions, a thought exercise was carried out to systematically explore the number of possible complexations from KIT, GAB1 and their shared binding partners. The models will demonstrate how SH2-peptide interactions may result in diverse and intricate behaviours and cellular response. It is important to note that only the simplest scenario was assumed, and as a proof of concept, the model only considered KIT, GAB1 and PI3K regulatory subunits. The reality is expected to involve more proteins with more types of alternations (e.g. other post-translational modifications) and complexations (e.g. SH3 to proline-rich peptide binding), which would result in an even higher level of complexity. Models proposed in Figure 4.9 and Figure 4.10 explore different complexation possibilities, providing a rational perspective into understanding the IDR-mediating PPI mechanisms.

The proposed models shown in Figure 4.9 considered mutually exclusive and mutually inclusive SH2 bindings in PI3K regulatory subunits. Like the other KIT and GAB1 shared binding partners, PI3K regulatory subunits have tandem SH2 domains (N-SH2 and C-SH2). Due to the high sequence conservation between SH2 domains in PIK3R1, PIK3R2, and PIK3R3, pY motifs that bind to PIK3R1 are assumed to interact with PIK3R2 and PIK3R3. While the models explore all the possible PPI complexations between the SH2 domains in PI3K regulatory subunits and the corresponding pY peptide motifs in KIT and GAB1, it is important to note that our models do not reflect the mechanistic reality. The models also assume that all SH2 domains have an equal binding affinity towards the pY peptide motifs. Figure 4.9A demonstrates a mutually exclusive SH2 binding in PI3K where only one of the SH2 domain is bound to a pY peptide motif. With four pY peptide motifs (one from KIT and three from GAB1), mutually exclusive SH2 binding will result in eight possible complexes. Figure 4.9B demonstrates a mutually inclusive SH2 binding where both SH2 domains are bound to a pY motif, resulting in 12 possible PPI complexes.



Figure 4.9: A schematic illustrating two proposed protein-protein interaction models of KIT, GAB1 and PI3K subunits. The N-SH2 and C-SH2 domain in PI3K regulatory subunits are shown in blue boxes. The pY peptide motifs from KIT and GAB1 are shown in orange boxes. It was assumed that all SH2 domains have an equal binding affinity towards the pY peptide motifs in KIT and GAB1. (A) Assuming only one SH2 domain from PI3K is bound to a pY motif from KIT or GAB1. N-SH2 binding and C-SH2 binding are mutually exclusive. This results in 8 possible complexations. (B) Assuming both C-SH2 and N-SH2 from PI3K are bound to one of the pY peptide motifs from KIT or GAB1. N-SH2 binding and C-SH2 binding and C-SH2 binding are mutually inclusive. This results in 12 possible complexations.

Prior research has proposed a three-state equilibrium binding model for other proteins with tandem SH2 domains [371, 372]. Figure 4.10 shows a model adopted from this concept for the KIT-PI3K and GAB1-PI3K PPIs. **State one** represents an inactive PI3K kinase with both N-SH2 and C-SH2 in the regulatory subunits bound to the catalytic subunit. The intermolecular binding between the regulatory subunits and catalytic subunits provides an inhibitory effect on PI3K's catalytic activities. Importantly, at this state, N-SH2 and C-SH2 are not bound to any pY peptide motifs from their binding partners. **State two** represents a semi-active form of PI3K, with one of the SH2 domain binds to a pY peptide motif. The SH2-peptide interaction relaxes the inhibitory effect. The N-SH2 binding and C-SH2 binding are mutually exclusive in

this stage. In **State three**, the inhibitory effect on its catalytic subunit is completely released, with both SH2 domains binding to a pY peptide motif. In this stage, the N-SH2 and C-SH2 binding are mutually inclusive.

While some suggested that N-SH2 acts as a primary regulator [373], others reported that C-SH2 binding initiates the kinase activities [374]. It is inconclusive whether N-SH2 or C-SH2 binds to the peptide to relax the inhibitory effect first. Nevertheless, this three-state regulatory mechanism shows an additional complexity to the PPI model and demonstrate how SH2-peptide PPIs could result in a number of different mechanisms and potentially regulating a range of biological responses.



Figure 4.10: This schematic illustrates a three-state equilibrium SH2 binding model in PI3K, adopted from previous studies [371, 372]. Blue boxes represent SH2 domains in PI3K regulatory subunits; green boxes represent PI3K catalytic domains; yellow boxes with thick black lines represent phosphorylated tyrosine peptide motifs; the red start represents activating kinase. In state one, both N-SH2 and C-SH2 in the regulatory subunits are bound to the catalytic subunit, inhibiting the kinase activities. In state two, one of the SH2 domains is bound to a pY motif, and the other SH2 is still bound to the catalytic domain. It is suggested that the kinase remains inactive at this state. In state three, both SH2 domains are bound to peptides, releasing the inhibition effect on the catalytic subunits. Kinase is activated at this state.

## 4.6 Key findings

- The PPI networks developed in this chapter demonstrated that ERBB2 and ERBB3 play an important role in bridging the LCK and STAT1 PPI network. However, previous studies have suggested that neither ERBB2 or ERBB3 is a suitable drug target for Basallike breast cancer.
- In the LCK network, KIT and GAB1 were the hubs and they shared a number of proteins binding partners, including PI3K regulatory subunits, RASR1, PTPN11 and PLCG1.
- The interactions between KIT, GAB1 and their binding partners were predicted to be mediated by phosphorylated tyrosine peptides and SH2 domains.

## Chapter 5

# **Structural Characterisation**

## 5.1 Introduction

The previous chapter identified a number of critical PPIs that can potentially be targeted as a drug target in Basal-like breast cancer. To examine whether these PPIs are mediated by the putative binding sites revealed in Chapter 4, molecular docking was performed in this chapter. Specifically, the KIT and GAB1 peptide sequences that contain SH2-binding SLiMs were docked on the N-SH2 and C-SH2 domains in the three PI3K regulatory subunits (PIK3R1, PIK3R2 and PIK3R3). The results presented in this chapter will provide an indication of (i) whether these sequences are likely to bind to the SH2 domain in PI3Ks, and (ii) if an interaction is observed, how the peptide (from the binding sequence) is bound to the SH2 domain. While this chapter only explored 6 of out the 12 critical PPIs, the strategy described in this chapter (e.g. KIT-PTPN11).

#### 5.1.1 Diversity of protein-protein interactions

Protein-protein interactions (PPIs) are structurally and functionally diverse. They differ based on their composition, affinity and the nature of association. *In vivo*, many factors could affect the interactions between proteins, including protein's cellular localisation, concentration and the local environment. These factors control the composition and oligomeric form of protein complexes, regulating a wide range of biological functions and activities. PPIs can be classified in different ways, and some of the common PPI types have been discussed by Nooren and Thornton [375]. Specifically, based on PPIs' interaction surface, a protein complex can be homo-oligomeric or hetero-oligomeric; based on their stability, a PPI can be obligate or nonobligate; and finally based on their persistence, a PPI can interact in a "transient" or a "stable" fashion [376].

#### Homo-oligomers and hetero-oligomers

Homo-oligomers are protein complexes constituted by identical chains; whereas hetero-oligomers are made up by non-identical chains. Homo-oligomers can be further categorised as isologues or heterologues based on binding similarity and differences respectively [377]. An isologous assembly involves binding between the same surface from both monomers. In contrast, heterologous associations are defined when the PPI occurs between different interfaces. Compared to the association between isologues where further oligomerisation can only occur between a new pair of interfaces, assemblies between heterologous are more versatile and diverse and have more potential for aggregating. Indeed, homo-oligomers and hetero-oligomers are often involved in distinct biological processes. For example, homo-oligomers are often seen in enzymes, carrier proteins or transcriptional regulatory protein machinery [378], whereas hetero-oligomers are often involved in regulating distinct and differential cellular functions, e.g. intracellular signalling [379].

#### Obligate and non-obligate complexes

Apart from their composition, protein complexes can also be classified as obligate or nonobligate PPIs. Obligate PPIs are made up of proteins that are not, or rarely, exist as a stable protein alone *in vivo*. They generally resemble a large globular structure with a large area of hydrophobic interface. Obligate PPIs ensure protein domain complexes occurs to form functional protein complexes, on disassociation, the functionality of lost [380]. For instance, many DNA binding proteins, such as Ku or Arc proteins, each function as obligate homodimers. In contrast, non-obligate PPI complexes involve complexation between protein domains that can exist independently *in vivo*. In comparison to obligate PPIs, non-obligate PPI complexes generally have smaller and less hydrophobic interfaces [380–382]. Many intracellular signalling PPI complexes are non-obligate, allowing an exchange of protein domain partners. For example, Ras proteins and G proteins interchangeably form non-obligate PPI complexes with GTPase activating proteins (GAPs).

#### Transient and permanent complexes

Based on the lifetime of a protein complex, a PPIs can be transient or permanent. Transient PPIs involves proteins association and disassociation temporarily *in vivo* and are usually reversible [383, 384]. Such complexations are often seen between signalling proteins and are typically specific to certain cellular contexts (e.g. cell type, cell cycle stage, localisation etc.), representing important biological regulators. In contrast are permanent PPIs which are generally very stable and often exist in their complex form and not reversible. One example of the permanent protein complex is IL-5 cytokine dimer (PDB ID: 3b5k) [385], which is a permanent protein-protein interaction. Functionally or structurally obligate PPIs are typically permanent [375], whereas non-obligate PPIs are predominantly transient [386].

It is important to note that while a given PPI may be a combination of these three PPI types, many PPIs do not fall into distinct categories. Moreover, the characteristics of PPIs often depends on physiological, ionic or redox conditions. For example, the stability of a PPI might be very different *in vivo* and under certain intracellular conditions or compartments. Measuring the affinity and dynamics of PPIs between endogenous proteins are not always feasible, therefore the cellular localisation and functionality of the proteins involved can help to suggest the type of interaction. For example, interactions between signalling proteins are expected to be nonobligate and transient, since the process of signal transduction requires frequent associations and dissociations with discrimination to activate distinct pathway and cascades.

#### 5.1.2 Intrinsically disorder regions in signalling proteins

Traditionally, protein function is assumed to be dependent on tertiary structure and PPIs were thought to be constrained by the surface of structured domains. For this reason, the functionality of intrinsically disordered regions (IDRs) in proteins has been overlooked. IDRs are considered to be protein regions that lack defined three-dimensional structures under intracellular conditions [387–389]. Recent research had revealed that IDRs constitute up to 40% of the residues in eukaryotic proteomes and are particularly enriched in highly connected hub proteins [390–393]. This observation is important to this project. It is now recognised that IDRs are involved in many biological processes, particularly in signal transduction and transcriptional regulation [394]. Their flexibility allows them to adapt to different conformational requirements and enables them to bind to multiple binding partners. Furthermore, IDR-mediated PPIs are typically transient, specific, and low-affinity [395]. It has been suggested that the structural and biochemical properties of IDRs allow them to constrain or control specific interactions or regulatory events spatially and temporally. This characteristic could be an adaptation of signalling protein evolution, where frequent and rapid associations and disassociations are required [396–398]. Compared to structured signalling proteins, those with IDRs have a larger interaction surface, higher degree of flexibility and more accessible sites for post-translational modification, allowing binding to different target proteins on different occasions. While it has been shown that IDRs play an essential role in constraining interactions between signalling proteins, IDR-mediated PPIs have previously been considered to be "undruggable" due to the lack of available structural insight. However, the draggability of IDR-mediating PPIs have been reconsidered in the last decade, and recent drug discovery has successfully developed peptidomimetics or small molecule-based inhibitors for IDR-mediating PPIs, such as Bcl-xL and BAK [399, 400], MDM2 and p53 [284], interleukin (IL)-2 receptor  $\alpha$ and IL-2 [401, 402] etc. These discoveries have encouraged the community to explore more IDRmediated PPIs, including SH2-peptide PPIs, that have been previously neglected and considered undruggable.

#### 5.1.3 SH2 domains

Src Homology 2 (SH2) domain is a structurally conserved protein domain that is well-characterised in many signalling proteins. It was first discovered by Tony Pawson and colleagues while they were studying the v-Fps/Fes oncoprotein [403]. They identified a non-catalytic region of a protein that shared sequence homology to Src-related kinases in numerous eukaryotic organisms [403, 404]. This conserved region, consists of approximately 100 residues, and was subsequently named the SH2 domain. Since then, SH2 domains have also been found in several other catalytic and non-catalytic proteins, including adaptors and signalling scaffolds [405, 406].

Functionally, SH2 domains play an important role in phospho-tyrosine signalling. They recognise phosphorylated-tyrosine (pY)-containing motifs within their target proteins, coordinating cellular signal transduction events immediate downstream of receptor tyrosine kinases (RTKs), adaptors, and scaffolds. They function closely with RTKs and protein tyrosine phosphatases to direct and control the spatial and temporal organisation of phosphorylated-tyrosine signalling. Through this mechanism, SH2 domains regulate many aspects of cell communications [278, 407]. They have also been shown to play a critical role in pathological conditions, including breast cancers [408], diabetes, and immunodeficiencies [409, 410]. The first SH2 crystal structure, solved by Kuriyan *et al.* [411, 412], revealed a central antiparallel  $\beta$ -sheet flanked by two  $\alpha$ -helices. Several key elements of SH2 domain have been shown to be critical for pY peptide binding. Some of these features were common to most human SH2 domains, such as a critical arginine in the positively charged pocket on the  $\beta$ 2 strand; and some were specific to a certain SH2 domain in a particular protein, providing ligand selectively and signal complexity.

While more than 300 SH2 domains have been solved and structurally characterised by X-ray crystallography or Nuclear Magnetic Resonance (NMR), they only account for about half of the known SH2 domains in the human proteome [409]. In this chapter, both homology modelling and molecular docking tools were used to explore the experimentally unsolved SH2-peptide complexations between KIT, GAB1 and PI3K regulatory subunits.

## 5.1.4 Aims

The work presented in this chapter aimed to address the following:

- to examine whether the putative peptide regions from KIT and GAB1 interact with the SH2 domains in PI3K regulatory subunits;
- if interactions are observed, do the peptides have preferences for N-SH2 or C-SH2 domains?
- to reveal and characterise the predominant binding modes for each of the PI3K SH2 complexations.

#### 5.1.5 Overview

The strategy used in this chapter is described in Figure 5.1. In the data collection and preparation step, relevant crystal structures were identified and collected from Protein Data Bank. If crystal structures are available for the SH2, the structure is directly used in docking. If no crystal structure is available for an SH2 domain, its three-dimensional structure was constructed by homology modelling. For the peptide, if it has been previously resolved with a SH2 domain, then peptide was extracted from the PDB file and used directly in the docking. If the sequence (short oligopeptide) has not been complexed with SH2 domain in a resolved structure before, then it was built in PyMoL. Once the peptide and SH2 domain structures are prepared, a peptide is placed on the SH2 binding site, as a proxy for the phosphorylated tyrosine binding sequences. Peptide docking protocol from Rosetta was then performed. The protocol has a preliminary energy minimising step called *prepacking*, which was followed by the docking protocol that generates an ensemble of 1000 peptide-SH2 conformations. The docking outcomes can be then analysed and interpreted using R and molecular visualisation tools.



Figure 5.1: The workflow of the peptide docking presented in this chapter. The workflow has two main steps, with the first step being data collection and preparation, and the second step being the peptide docking. In the first step, the availability of relevant PDB crystal structures was explored, and input models for docking were prepared. In the second step, the peptide docking protocol from Rosetta was used for peptide-SH2 conformational sampling. In each docking simulation, an ensemble of 1000 models was generated.

In this chapter, N-SH2 and C-SH2 domain in PIK3R1 are referred to as R1N and R1C; similarity, N-SH2 and C-SH2 domain in PIK3R2 and PIK3R3 are referred to as R2N, R2C, R3N and R3C, respectively. Oligopeptide are considered as proxy binding sequences, representing a complementary strand of a binary partner to the SH2 domain. Moreover, peptide residues are described in their position in relation to the phosphorylated tyrosine (pY), with residues N-terminal to the pY being referred to as pY+1, pY+2, pY+3, and so on, and residues C-terminal to the pY being referred to as pY+1, pY+2, pY+3, and so on.

## 5.2 Data collection and docking protocols

In this chapter, three phosphorylated tyrosine (pY)-containing peptides were docked onto the N-SH2 and C-SH2 domains in PI3K regulatory subunits (PIK3R1, PIK3R2, PIK3R3). Data preparation and docking protocol used in this chapter will be explained in detail in the following subsections.

#### 5.2.1 Available PDB crystal structures

In order to gain structural insight for the SH2-peptide complexations in PI3K regulatory subunits, relevant high-resolution (<2.5 Å) crystal complexes were collected from Protein Data Bank (PDB), an archive of experimentally determined 3-dimensional structures (typically crystals or NMR ensembles) of biological macromolecules. Figure 5.2 demonstrates the process of identifying and collecting the relevant crystal structures. Three human PI3K regulatory subunits, PIK3R1, PIK3R2 and PIK3R3, were initially searched on PDB (https://www.rcsb.org) and three criteria were applied to filter the undesired crystal structures. Each of the selected crystal complexes contained at least one SH2 domain from PI3K regulatory subunits and a corresponding pY-peptide ligand. Table 5.1 shows the detail of selecting the crystal structures. Of note, no crystal structure was found for PIK3R2 or PIK3R3.



Figure 5.2: Process for identifying relevant pre-solved crystal structures from the **PDB**. Three selection criteria were applied. The first criterion was whether the PDB structure contains at least one SH2 domain from PI3K. The second criterion was whether the SH2 domain in the complex is complexed with a phosphorylated tyrosine (pY) peptide. The third criterion was whether the resolution of the PDB 2.5 Å or better. Six PDB structures were identified for PIK3R1, while no crystal structures were available for PIK3R2 and PIK3R3.
Protein	PDB ID	Peptide	Resolution (Å)	SH2 domain (N-/C-/Both)
PIK3R1	1H9O	pY VPML	1.79	C-SH2
PIK3R1	1PIC	pY VPML	$\operatorname{Ensembles}(\operatorname{NMR})$	C-SH2
PIK3R1	2IUH	TNE pY MDMK	2.00	N-SH2
PIK3R1	2IUI	SID pY VPMLMDMK	2.40	N-SH2
PIK3R1	$5 \mathrm{GJI}$	SD pY MNMTP	0.90	N-SH2
PIK3R1	5AUL	SD pY MNMTP	1.10	C-SH2

Tabl	e 5.1:	Details	of the	e relevant	crystal	structures	$\operatorname{col}$	lected	from	PDE	3
------	--------	---------	--------	------------	---------	------------	----------------------	--------	------	-----	---

In this work, these resolved crystal structures of PIK3R1 SH2-peptide interactions were used to determine the optimised length and orientation of the SH2 binding pY peptides.

### 5.2.2 Data preparation

In this chapter, one pY peptide from KIT and three from GAB1 were docked onto the N-SH2 and C-SH2 domains in PI3K regulatory subunits, resulting in 24 docking experiments (see Figure 5.3). Each of experiment required an SH2 domain and a corresponding pY peptide ligand. SH2 domains were extracted from the relevant crystal coordinates. If no crystal structure was available, homology modelling was performed using Phyre2 (www.sbg.bio.ic.ac.uk/phyre2/) to predict a 3D structure of the SH2 domains.

Previous studies have revealed several 14-mer SH2-binding sequences from KIT and GAB1 (shown in Chapter 4), the crystals identified from the PDB showed that most residues contributing to the binding (either forming favourable contacts or having steric or functional complementary) are close to the pY. To reduce the computational cost and ensure the simulations are manageable, shorter amino acid sequences (eight for KIT peptide and seven for GAB1 peptides) were used in the simulations. Crystal coordinate also provided an indication of the optimal dihedral angles for binding to the SH2 domain. Peptide ligands were created in PyMoL with their geometries adjusted to optimal orientations. The SH2 domain and the pY-peptides were superimposed on a PDB template (PDB 5GJI for N-SH2 models and PDB 5AUL for C-SH2 models). Templates are defined as having the best resolution or most relevant PDB available for the given PPI.



Figure 5.3: A list of pY peptides (magenta) and SH2 domains (green) considered in this chapter. The three peptides from KIT and GAB1 have been previously predicted to bind to PIK3R1 (see Chapter 4), however, whether they do in fact bind to PIK3R2 and PIK3R3 remains unclear. Each of the peptides were docked to the SH2 domains in PIK3R1, PIK3R2 and PIK3R3, giving rise to 24 docking simulations.

# 5.2.3 Preliminary stage

In this chapter, FlexPepDock docking protocol from Rosetta (v3.10) was performed, in which a preliminary stage calls *prepacking* is required. This stage removes internal clashes in the SH2 domain and peptide by the initial packing of the side chains. Internal clashes removed at this stage are unlikely to disturb any inter-molecular interactions [413]. This stage involves conformational sampling of the side chains to determine the best rotamer combinations for both SH2 domain and peptide while the backbones remain fixed. The best (lowest total energy score) output structure from this stage was used as the initial geometry for the docking protocol.

## 5.2.4 Peptide docking

In this work, peptide docking was performed for 24 SH2-peptide complexes using Rosetta (v3.10), with 1000 models generated per docking simulation. The starting structure is obtained from the preliminary prepacking stage. As the pY peptides were posed and placed close to the SH2 binding pockets prior docking, it was assumed that the SH2-peptide complexes were approximation of a larger complexation with complementary sequences. Instead of running the blind global docking protocol intended for cases with no available information about the peptide backbone conformation, the Rosetta FlexPepDock refinement protocol was used to optimise the pY peptide backbone iteratively and its rigid-body orientation relative to the SH2 domain. This generates high-resolution models with side chains of the binding motifs modelled at an atomic accuracy. To emphasise perturbations in the binding site while preventing the separations between the peptide and the domain during an energy minimisation, the refinement protocol starts with increasing and decreasing the weights of the repulsion and attraction van der Waals respectively. The flags used for the refinement protocol are described in Appendix I.

#### 5.2.5 Conformational energy distribution

Apart from the 1000 ensembles, each docking simulation also produces an output scoring file containing several scoring functions that predict the binding affinity between the SH2 domain and the pY peptide ligand for a given geometry. In this chapter, the *interface score* was plotted against *interfacial all atom Root Mean Squared Deviation (RMSD)* for all ensembles from each of the simulations. The *interface score* describes the free energy contributed by interface residues of each given complex, with negative values representing favourable contacts. Of note, energy scores produced by Rosetta cannot be directly converted to physical energy units (e.g. kcal/mol); instead, they are represented using Rosetta Energy Unit (REU) in Rosetta (a proxy for free energy). On the other hand, the *interfacial all atom RMSD* measures the geometric changes of the interfacial residues between starting and final structure, where interfacial residues include any residue whose C $\beta$  atom (C $\alpha$  for glycine) is within 8Å of a C $\beta$  atom on the binding partner.

Interface score vs interfacial all atom RMSD scatter plots were generated and shown in this chapter. These provide an indication of how the geometry of the interface affects the free energy of binding. Histograms were added to the x and y axis to show the marginal distribution of the ensemble of the 1000 models. Scatter plots were generated using R (v3.6.1) and R/ggplot2 (v3.20).

#### 5.2.6 Focussed density contour plots

Highly populated areas were often found in the *interface score* vs *interfacial all atom RMSD* scatter plots. To better interpret these high density areas, outliners were excluded, and data of the top 90% of both distributions (in x and y) were used to generate focused density contour plots. Such plots represent density distribution of the data, indicating the height and gradient of the contours of the populated regions. Peaks with populated regions are assumed to represent different binding modes, and representative complexes were subsequently identified for further analysis. The density contour plots were generated using R (v3.6.1) and R/ggplot2 (v3.20).

#### 5.2.7 Representative structure evaluation

In order to select the representative models from an ensemble, x and y values that represent each high density region were identified. Geometries that satisfy the x and y criteria were selected and referred to as representative structures. These representatives were then visualised and analysed using PyMol (v2.3.3). If representatives for a given high density region represents more than one binding modes, then these are further classified into subgroups (A1, A2 etc.). Also hydrogen bonds identified in this chapter had atom to atom (donor-acceptor) distance of 3 Å, considered as strong-moderate hydrogen bonds [414]. In the following sections, peptides are representing the SLiMs that potentially are bound to SH2 domains.

# 5.3 An ensemble of KIT-PI3K interactions

The 8-mer pY peptide TENpYMDMK from KIT AA718 to AA725 was docked onto both N-SH2 and C-SH2 domains in PIK3R1, PIK3R2 and PIK3R3. While previous studies only demonstrated KIT-PIK3R1 and KIT-PIK3R2 PPIs, it was assumed that due to the high sequence similarity of the SH2 domains among the PI3K regulatory subunits (shown in Chapter 4), this KIT pY peptide sequence may have the potential to complex with the SH2 domains in PIK3R3.

The energy score plots of the ensembles for the six KIT to PI3K regulatory subunits complexations are shown in Figure 5.4. Table 5.2 summarises the x axis (*interfacial all atom RMSD*) and y axis (*interface score*) range of the top 90% of the data for the distributions in both dimension for each docking simulation. Models outside these ranges were considered as outliners and excluded from the density contour plots. KIT-R1N PPI can be seen to deviate less (within 2.0 Å) compared to the other five KIT-PI3K PPIs (within 4.0-5.0 Å), suggesting the starting structure of KIT-R1N PPI was close to the favourable binding geometries. Moreover, while the *interface score* for the majority of the models falls below -25 REU, some models from the KIT-R1C simulation have higher interface score (between -25 and 0), indicating less favourable complexations.

Density contour plots revealed highly populated areas. In general, a positive correlation can be seen between the *interface score* and the *interface score* for the N-SH2 complexations. However, this correlation is less strong in the C-SH2 PPIs. This suggests that the free energy of binding is affected by the interface geometries more in the N-SH2 PPIs and less in the C-SH2 PPIs.

In N-SH2 PPIs (Figure 5.4 left), PIK3R1 and PIK3R2 complexations have two distinct binding modes A and B whereas PIK3R3 only has a single broad binding mode. In the C-SH2 PPIs (Figure 5.5 right), the density plot for the PIK3R1 PPI revealed four high density regions (A, B, C, D), however only region A and region B present distinct binding modes. PIK3R2 and PIK3R3 C-SH2 domain PPI also have two high density regions each, all presenting distinct peptide geometries.



Figure 5.4: Scatter and density contour plots from KIT 718-725 to PI3K SH2 domain simulations. The scatter plots demonstrate the relationship between *interfacial all atom* RMSD and *interface score* for each of the simulation. Each dot in the scatter plot represents a model from the ensemble. A marginal distribution of the two variables is shown on the x and y axes. Density contour plots show 90% of the data presented in the respective scatter plot. Contour lines connect points of equal density in this graph.

Table 5.2: The values for 90% of the x and y distributions for each KIT 718-725 to PI3K SH2 domain simulation. Models within these ranges were included in the density contour plots, models outside these values were treated as outliners and eliminated from the analysis. The *interfacial all atom RMSD* measures the geometric changes of the interfacial residues between starting and final structure. The *interface score* describes the free energy contributed by interface residues of each given complex, with negative values representing favourable contacts.

	Interfacial all atom RMSD (Å)	Interface score
KIT-PIK3R1 N-SH2	x<2.0	y<-35
KIT-PIK3R1 C-SH2	x<4.0	y<0
KIT-PIK3R2 N-SH2	x<4.5	y<-25
KIT-PIK3R2 C-SH2	x<5.0	y<-20
KIT-PIK3R3 N-SH2	x<4.5	y<-20
KIT-PIK3R3 C-SH2	x<4.5	y<-20

Figure 5.5 shows the representative models for the high density areas identified from the KIT-R1N (top panel) and KIT-R1C (bottom panel). The representative models from KIT-R1N PPI show that while the side chain conformation of the pY and the C-terminal Lysine are different in A and B, the two peptides interact with similar residues from the N-SH2 domain. Residues from N-SH2 domain that contribute to the peptide binding includes 22N, 40S, 56N, 57N, 58K, 59L and 96N. While this highlights the involvement of these residues for recognising this particular pY peptide, it was surprising that most of these residues (apart from 40S) are interacting with the peptide backbone and not to the pY.

The bottom panel of Figure 5.5 shows the representative models identified from the KIT-R1C simulation, of which two distinct binding modes (B1 and B2) were revealed from region B. In structure A, the pY fits into a positively charged pocket with the phosphate group interacting with 38R, 40S and 41S. However, these interactions were absent in structure B1 and B2. In terms of the residues N-terminal to the pY, peptide backbone orientations are similar in structure A and structure B1, in which they interact with 24E, 57K, and 58H. Whereas in structure B2, the residues N-terminal to the pY are parallel to the beta-C strand and interact with several residues from beta-C, including 55E, 56V, 57K, 58H.



Figure 5.5: Representative structures revealed from KIT 718-725 to SH2 domains in **PIK3R1.** Two high density regions were identified from R1N (top), from which each revealed a distinct binding mode. Two high density regions were identified from R1C simulation (bottom), in which one and two binding modes were revealed from structure A and B, respectively. Peptides are shown in magenta sticks, and SH2 domains are shown in green cartoon. N- and C-ends of the peptides are labelled.

Figure 5.6 demonstrates the representative models for the KIT-R2N (top) and KIT-R2C (bottom) PPIs. Two high density regions (A and B) were revealed from the KIT-R2N simulations. In which two peptide poses (A1 and A2) were identified from region A, and one was identified from region B. Peptide pose A1 and A2 have distinct N-tails (pY-2, pY-3) and C-tails (pY+3, pY+4), whereas the geometries of the central peptides (pY-1 to pY+2) were very similar. Comparing to A1 and A2, structure B has very different side chain coordinates, especially the side chain of pY. Importantly, all three peptides interact with similar residues from the N-SH2 domain in PIK3R2. In particular, residue 45N, 46N, 47K and 48L from the N-SH2 domain interact with the peptide residue pY-3, pY-2, pY-1 and pY, respectively. This suggests that N-terminal residues are more involved in the SH2-peptide complexation than C-terminal residues in this particular PPI.

Similarly, two high density regions (A and B) were also revealed from the KIT-R2C simulation, with each region revealing one peptide pose (Figure 5.6 bottom). While the two poses have similar backbone orientations, their side chain orientations are distinct. In particular, their pY side chains are almost 180 degrees apart from each other. The distinct side chain orientation means that they interact with different residues from the C-SH2 domain. For example, pY in structure A interact with 8R and 26R, whereas pY in structure B interact with 51R. Moreover, the side chain of N-terminal residues in structure A are pointing away from the SH2 domain, too far to form covalent bonds with the C-SH2 domain. In contrast, the side chain of N-terminal residues in structure B were pointing towards the SH2 domain, which allows them to form hydrogen bonds with the residues from the alpha-A helix or beta-C strand of the C-SH2 domain.



Figure 5.6: Representative structures revealed from KIT 718-725 to SH2 domains in PIK3R2. Two high density regions were identified from R2N (top), in which one and two binding modes were revealed from structure A and B, respectively. Two high density regions were identified from R2C simulation (bottom), and each revealed a distinct binding mode. Peptides are shown in magenta sticks, and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

Figure 5.7 shows the representative models from KIT-R3N (top) and KIT-R3C (bottom) simulations. Only one high density region was revealed from KIT-R3N docking simulation, and the representative structure shows that the N-terminal residues interact with several residues from the beta-C strand, including 45N, 46N, 47K, 48L, 49I, 50K. This suggests that this complexation is driven by the specific molecular recognition between the beta-C strand and the N-tail peptide.

On the other hand, two representative structures were identified from the high density regions from the KIT to R3C docking simulation (Figure 5.7 bottom). In structure A, the peptide interacts with not only residues from beta-C strand but also residues from the loop5, loop6, alpha-A and alpha-C. In particular, residues from beta-C strand bind to the pY-3 to pY residues, stabilising the N-terminal of the peptide; whereas 59A from loop5 and 83H, 84N from loop6 interact with the pY+2 and pY+4 positions, stabilising the C-terminal of the peptide. Moreover, 8R from the alpha-A interacts with the phosphate from the pY. 8R remains a critical phosphate binding residue in structure B.

Different from structure A, the peptide in structure B only interacts with one residue, 46H, from the beta-C strand. The loss of covalent bonds between the peptide and beta-C strand is caused by the change of the peptide backbone orientation, particularly the residues at the N-terminal (pY-3 to pY). Moreover, in structure B, the N-terminal peptide is bound to 12E from alpha-A and 43E from the end of loop4 and the beginning of beta-C. Residue 12E and 43E also interact with the peptide in structure A, making them general peptide binding residues and not specific to this binding mode. Another critical residue is 8R which forms a hydrogen bond with the phosphate in both structures. In structure B, the phosphate group forms an extra hydrogen bond with 28S from loop3. The geometry of C-terminal peptide (pY+1 to pY+4) in structure B is very different from structure A. The C-terminal peptides from the two structures bind to very different regions of the SH2 domain. In particular, residue pY+3 from structure B form a hydrogen bond with 75Y from alpha-B, whereas residue pY+3 and pY+4 from structure A form hydrogen bonds with residues from alpha-C, loop5 and loop6 respectively.



Figure 5.7: Representative structures revealed from KIT 718-725 to SH2 domains in PIK3R3. Only one high density region was identified from R3N (top). Two high density regions were identified from R3C simulation (bottom), and each of them revealed a distinct binding mode. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

In summary, the above analyses show that N-SH2 binding is more favourable (lower *inter-face score*) than C-SH2 binding in PIK3R1 and PIK3R3 PPIs, while C-SH2 binding is more favourable than N-SH2 binding in PIK3R2 PPIs. This suggests that if both N-SH2 and C-SH2 are free (not occupied and without steric hindrance), KIT peptide is more likely to bind to N-SH2 in PIK3R1; C-SH2 in PIK3R2 and N-SH2 of PIK3R3 than the other SH2 domain in the protein. However, the differences of *interface score* between the N- and C-SH2 binding in PIK3R2 and PIK3R3 are only 1 and 0.5 respectively, and such complexations are also expected to be influenced by other factors such as intracellular microenvironment and the surrounding molecules (e.g. water, charged atoms). Moreover, while PIK3R1 PPI has a generally higher *interface score*, PIK3R2 and PIK3R3 could be as likely as PIK3R2 to interact with the pY peptide from KIT.

# 5.4 An ensemble of GAB1-PI3K interactions

As previously shown in Figure 5.3, GAB1 440-453, 465-478 and 582-592 were reported to complex with PIK3R1 by other studies. Similar to the assumption made before, due to the high sequence similarity of SH2 domains among PI3Ks, the three stretches of sequences from GAB1 were assumed to have the potential to interact with PIK3R2 and PIK3R3. Three pY peptides from GAB1 were docked on to the SH2 domains of the PI3K regulatory subunits. The results are shown in the following sections.

#### 5.4.1 GAB1 444-450 to PI3K regulatory subunits

The first peptide sequence DENpYVPM represents amino acids 444 to 450 in GAB1. Figure 5.8 showed the energy plots of the ensembles for the GAB1 444-450 to PI3K PPIs. Table 5.3 summarises the x (*interfacial all atom RMSD*) and y (*interface score*) cut off for eliminating the outliners. The summary shows that models for GAB1 440-450 to R2C PPI deviate more (within 6.0 Å) compared to other five PPIs (within 3.0-4.0 Å), indicating that the starting structure of GAB1 440-450 to R2C was further from to the favourable binding geometries than the other five complexes. In terms of binding energy, while the majority of the models have an *interface score* below -10 REU, some models from the GAB1 444-450 to the R1N and R1C simulations have higher interface score (between -10 and 0) and this could indicate less favourable complexation or increased instability with the SH2 surface.

The density contour plots reveals highly populated areas. In N-SH2 PPIs (Figure 5.8 5.8 left), four high density regions were revealed from PIK3R1 PPI and only a single high density region was revealed from PIK3R2 and PIK3R3 PPIs. In the C-SH2 PPIs (Figure 5.8 right), the density plot for the PIK3R1 PPI revealed two distinct high density regions. Similarly, three and two high density regions were revealed from PIK3R2 and PIK3R2 and PIK3R3 PPIs, respectively.



Figure 5.8: Scatter and density contour plots from GAB1 444-450 to PI3K SH2 domain simulations. The scatter plots demonstrate the relationship between *interfacial all atom RMSD* and *interface score* for each of the simulations. Each dot in the scatter plot represents a model from the ensemble. A marginal distribution of the two variables is shown at the x and y axis. Density contour plots show 90% of the data presented in the corresponding scatter plot. Contour lines connect points of equal density in this graph.

Table 5.3: The values for 90% of the x and y distributions for each GAB1 444-450 to PI3K SH2 domain simulation. Models within these ranges were included in the density contour plots, models outside these values were treated as outliners and eliminated from the analysis. The *interfacial all atom RMSD* measures the geometric changes of the interfacial residues between starting and final structure. The *interface score* describes the free energy contributed by interface residues of each given complex, with negative values representing favourable contacts.

	Interfacial all atom RMSD (Å)	Interface score
GAB1 444-450 to R1 N-SH2	x<4.0	y<0
GAB1 444-450 to R1 C-SH2 $$	x<4.0	y<0
GAB1 444-450 to R2 N-SH2 $$	x<3.0	y<-10
GAB1 444-450 to R2 C-SH2 $$	x<6.0	y<-15
GAB1 444-450 to R3 N-SH2 $$	x<4.0	y<-15
GAB1 444-450 to R3 C-SH2 $$	x<6.0	y<-10

Figure 5.9 shows the representative structures for the high density regions from the GAB1 444-450 to PIK3R1 PPIs. While four high density regions were identified from the GAB1 444-450 to R1N complexation, their representative models all have similar peptide geometries, suggesting that the four high density regions are only reflecting the geometric differences in the SH2 domain and not in the pY peptide. The example structure shows that the N-terminal peptide (pY-3 to pY) is almost parallel to the beta-C strand. This parallel orientation allows the N-terminal residues to bind to residues from the beta-C strand, including 56N, 57N, and 59L. Moreover, the phosphate group on the pY interacts with 37R from beta-A, 40S from loop 40 and 48T from beta-B. Finally, the proline at pY+2 also represents a critical residue, as it provides backbone rigidity to the C-terminal peptide and its carbonyl group form a stabilising contact with 96N from loop 9.

Two representative models representing the high density regions from the GAB1 44-450 to R1C PPI are shown at the bottom of Figure 5.9. The peptide backbone conformations are similar in structure A and structure B, with bot N-terminal resides (pY-3 to pY-1) binding to the 24E and 56V and pY+2 proline binding to the 96N. However, the pY side chain orientations are very distinct in the two structures. The phosphate group from the structure A fits into a positively charged binding pocket and form several hydrogen bonds with the residues that make up the pocket (e.g. 20R, 38R, 40S and 41S). On the other hand, the phosphate group from structure B fits into a large cavity next to the pocket. However, the residues that made up of this cavity are not close to the pY enough to form strong (less than 2.5 Å) hydrogen bonds. This lack of interactions in the structure B complexes are reflected at the free energy of binding (figure 5.8 top right), where high density region B is 10 REU less negative (less stable) than high density region A.



Figure 5.9: Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R1. While four high density regions were identified from R1N (top), the representative structure identified from each of the regions all has similar binding pose. Two high density regions were identified from R1C simulation (bottom), and each of them revealed a distinct binding mode. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

Figure 5.10 shows the representative models for the high density regions from the GAB1 444-450 to PIK3R2 PPIs. While only one high density region was revealed from the GAB1 444-450 to R2N simulation, two slightly different peptide poses were revealed (A1 and A2). In both structures, the N-terminal peptide (pY-3 to pY) binds to residues in the beta-C strand, specifically with the amino group from pY binds to 48L. However, the orientation of the pY side chain is different in the two structures, pY in the structure A1 fits into a positively charged pocket that is made up of 8R and 29S; whereas the phosphate group in structure A2 binds to the positively charged 50R from the beta-C strand.

Regarding the GAB1 444-450 to R2C simulation, three representative models were identified. In structure A, the phosphate group on the pY fits into a positively charged pocket, and interact with 8R, 26R ,28S and 30Q. These interactions were lost in structure B and structure C, potentially resulting in higher *interface score* (Figure 5.8 middle right). Moreover, in structure A, the secondary carboxyl group from the 12E form a strong hydrogen bond (1.8 Å) with the amino group from the aspartic acid at the pY-3. This interaction is also seen in structure B and C, suggesting that 12E is a crucial residue for N-terminal peptide binding. Finally, while C-terminal peptides in all three structures have very few or no hydrogen bond interaction, the proline at pY+2 provides conformational constraints to the peptide backbone. This conformational restriction imposed by proline allows the side chain of valine and methionine at pY-1 and pY-3 to fit into two separate hydrophobic pockets.



Figure 5.10: Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R2. While only one high density region was identified from R2N (top), two distinct binding modes (A1 and A2) were revealed from the representative models. Three high density regions were identified from R2C simulation (bottom) and each of them revealed a distinct binding mode. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

Figure 5.11 shows the representative models for the high density regions from the GAB1 444-450 to PIK3R3 PPIs. Two peptide poses (A1 and A2) were identified from the high density region A from the GAB1 444-450 to R3N simulation. In both structures, the N-terminal peptide (pY-3 to pY-1) binds to the residues in the beta-C strand, including 45N, 46N and 48L. However, the orientations of the pY and pY+3 side chains are different in A1 and A2 peptides, with the pY in A1 facing towards the alpha-A helix and pY in A2 being parallel to the beta-C strand. Being parallel to the beta-C strand allows the phosphate group in A2 to form an extra hydrogen bond with 52Y from the beta-C strand.

While two high density regions (A and B) were identified from the GAB1 444-450 to R3C simulation, only region A revealed a distinct binding pose. The C-terminal peptide in structure A forms a number of hydrogen bonds with residues from the alpha-A helix and beta-C strand. Importantly, like other GAB1 to C-SH2 PPIs, the amino group in the pY binds to the 46H from the beta-C strand. On the other hand, no hydrogen bonds were formed between the SH2 domain and the N-terminal peptide. Instead, the two hydrophobic residues at pY+1 and pY+3 fit into the hydrophobic pockets on the SH2 domain surface, demonstrating hydrophobic contacts.



Figure 5.11: Representative structures revealed from GAB1 444-450 to SH2 domains in PIK3R3. While only one high density region was identified from R3N (top), two distinct binding modes (A1 and A2) were revealed from the representative models. Only one high density region was identified from R3C simulation (bottom). Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

### 5.4.2 GAB1 469-475 to PI3K regulatory subunits

Similarly, the second 7-mer peptide sequence EANpYVPM from GAB1 AA 469 to 475 was docked on to the N- and C-SH2 domains in PI3K regulatory subunits and the results are shown in the following sections. Figure 5.12 showed the energy plots of the ensembles for the GAB1 478-484 to PI3K PPIs. Table 5.4 summarises the values for 90% of the x (*interfacial all atom RMSD*) and y (*interface score*) distributions shown in the scatter plots. From the summary, it can be seen that while the majority of the models have an *interface score* below -10 REU, some models from the R1C and to R3C simulations have higher *interface scores* (between -10 to 0). This suggests that R1C and R3C PPIs are generally less stable/favourable and (higher  $K_{off}$ ) than the other PPIs. The negative (<0) *interface scores* of R1C and R3C still indicate that this pY peptide can bind to R1C and R3C SH2 domain more favourable (without the involvement of a third entity).

The density contour plots revealed high density regions from each docking experiment. In N-SH2 PPIs (Figure 5.12 left), two high density regions were revealed from PIK3R1 PPI whereas the PIK3R2 and the PIK3R3 simulations revealed a single high density region. In the C-SH2 PPIs (Figure 5.12 right), the density plot for the PIK3R1 PPI revealed three distinct regions, and the density plot for PIK3R2 and PIK3R3 PPIs revealed one high density region each. Representative models identified from each of the high density regions are shown and discussed below.



Figure 5.12: Scatter and density contour plots from GAB1 469-475 to PI3K SH2 domain simulations. The scatter plots reveal the relationship between *interfacial all atom* RMSD and *interface score* for each of the simulation. Each dot in the scatter plot represents a model from the ensemble. A marginal distribution of the two variables is shown at the x and y axis. Density contour plots show 90% of the data presented in the corresponding scatter plot. Contour lines connect points of equal density in this graph.

Table 5.4: The values for 90% of the x and y distributions for each GAB1 478-484 to PI3K SH2 domain simulation. Models within these ranges were included in the density contour plot, models outside these values were treated as outliners and eliminated from the analysis. The *interfacial all atom RMSD* measures the geometric changes of the interfacial residues between starting and final structure. The *interface score* describes the free energy contributed by interface residues of each given complex, with negative values representing favourable contacts.

	Interfacial all atom RMSD (Å)	Interface score
GAB1 478-484 to R1N	x<3.5	y<-10
GAB1 478-484 to R1C	x<4.5	y<0
GAB1 478-484 to R2N $$	x<3.0	y<-10
GAB1 478-484 to R2C $$	x<6.0	y<-17
GAB1 478-484 to R3N $$	x<4.0	y<-10
GAB1 478-484 to R3C	x<4.5	y<0

Figure 5.13 shows the representative models for R1N (top) and R1C (bottom) PPIs. Two high density regions (A, B) were identified from the R1N docking experiment and the corresponding representative models show that the peptides in both structures share a similar N-terminal orientation, with the glutamic acid at pY-3 and asparagine at pY-1 binding to the 57N and the backbone of pY binding to 59L. However, the orientation of the pY side chain is distinct in the two structures. Specifically, the pY side chain in structure A is perpendicular to the beta-C strand and binds to 40S and 19R; whereas the pY side chain in structure B is parallel to the beta-C strand and binds to 44H. This differences in pY side chain orientation could affect the interfacial binding energy and lead to the differences in *interface score* between region A and region B (Figure 5.12 top left).

Three high density regions (A, B and C) were identified from the GAB1 478-484 to R1C simulation, in which one binding mode was identified from region A, two binding modes were revealed from region B and one binding mode was region C. In structure A, N-terminal peptide (pY-3 to pY) is parallel to the beta-C strand and interacts with 55, 56V, 57L and 58H. While most of these interactions are lost in structure B1, B2 and C, 58H from beta-C strand interact with pY backbone in all four geometries, presenting a key contact for the complexation. Moreover, the proline at pY+2 interacts with 96N in both structure A and C, providing conformational constraints to the peptide backbone and force the two hydrophobic residues at pY+1 and pY+3into two hydrophobic pockets separately. These hydrophobic interactions between the SH2 surface and the C-terminal peptide are favourable contacts, contributing to the complexations. In structure A, B1 and B2, the phosphate group on the pY bind to several positively charged or polar residues, including 38R, 40S and 41S. Apart from the other interactions, 38R, 40S and 41S also form an interfacial cavity for pY binding, demonstrating a level of shape complementarity between the phosphate group and the SH2 domains. As previously shown in Figure 5.12, the interface score of structure C is less negative (less stable) than region A but more negative (more stable) than region B, the molecular features of these structures suggest that the loss of this pY binding pocket interaction in structure C seems to have less effect on the *interface score* than the backbone conformational constraints provided by the proline-96N interaction.



Figure 5.13: **Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R1.** Two high density regions (A and B) were revealed from the R1N (top) with each representing a distinct binding mode. Three high density regions (A, B, C) were identified from the R1C simulation (bottom). Of which region A and C represent a distinct binding mode and region B revealed two binding modes (B1 and B2). Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

Figure 5.14 shows the representative structures identified from R2N (top) and R2C (bottom) PPIs. In both simulations, only one structure was revealed. In the R2N representative structure, residues N-terminal to the pY binds to several residues from beta-C strand including 45N, 46N, 48L and 50K. In addition, a hydrogen bond is formed between the carboxyl group from the value at pY+1 and 84Y. This and the pY-48L interaction lock the peptide in an orientation which the hydrophobic value side chain can fit into a hydrophobic pocket.

The N-terminal peptide (from pY-3 to pY) from the R2C structure (Figure 5.14 bottom) is also in parallel to the beta-C strand, with the peptide binding to 43D, 44T and 46H. The phosphate group on the pY is also in parallel to the beta-C strand, however, the beta-C strand in the C-SH2 domain is a lot shorter than the beta-C in N-SH2 domain. Therefore, instead of binding to residues from the beta-C strand, the phosphate group in this structure binds to the 30Q from loop 3.



Structure A

Figure 5.14: Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R2. Only one high density region was revealed from R2N (top) and R2C simulation (bottom) each. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

Similarly, Figure 5.15 shows the representative structures from GAB1 478-484 to R3N (top) and R3C (bottom) PPIs. In which two distinct peptide poses were revealed from region A from the R3N simulation. In both A1 and A2, the N-terminal peptide interacts with a number of residues from the N-end of the beta-C strand. However, the pY side chain and the phosphate group attached to the pY have a distinct orientation in the two structures. Specifically, the pY side chain in structure A1 is pointing away from the beta-C strand and does not interact with any residues from the beta-C strand, whereas the pY side chain in structure A2 is in parallel to the beta-C strand and interact with the 50K and 52Y. As both structures were revealed from the same high density region and have a similar *interface score*, the differences in pY side chain conformations seem to have limited effect on the free energy of binding.

Two structures, A1 and A2, were revealed from the high density region A from the R3C docking simulation (Figure 5.15 bottom). While two small peaks were revealed from region A (yellow arrows in Figure 5.12 bottom right), it was hard to determine whether the A1 and A2 structures represent these two small peaks. In structure A1, the glutamic acid at pY-3 forms a hydrogen bond with the 43E and the pY backbone interacts with the 46H from the beta-C strand. These favourable contacts were also observed in structure A2. While N-terminal peptides have a similar orientation in the two structures, the pY side chain in A1 and A2 are distinctly different. Specifically, in structure A1, pY is facing away from the plane and pointing into a pocket of the SH2 domain, forming hydrogen bonds with 26R and 28S. However, in structure A2, the pY side chain is pointing towards the C-terminal end of the peptide and does not form hydrogen bonds with any residues from loop3.



Figure 5.15: Representative structures revealed from GAB1 469-475 to SH2 domains in PIK3R3. One high density region was identified from R3N simulation (top), and two binding modes (A1, A2) were revealed from the region. Similarly, one high density region was revealed from R3C simulation (bottom), revealing two binding modes (A1 and A2). Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the the peptides are labelled.

### 5.4.3 GAB1 586 to 592 to PI3K regulatory subunits

Finally, the third 7-mer peptide sequence EENpYVPM from GAB1 AA 586 to 592 was docked on to the SH2 domains in PI3K regulatory subunits. Figure 5.16 shows the energy plots of the ensembles for the GAB1 586 to 592 to PI3K docking simulations. While the previous docking results suggest that GAB1 444-450 and GAB1 478-484 interacts with both N-SH2 and C-SH2 domains in the PI3K regulatory subunits, docking of GAB1 486-592 to R1N and R2N failed. The unsuccessful docking could be caused by clashes or unfavourable contacts formed between the peptide and residues at the surface of R1N and R2N SH2 domains. Table 5.5 summarises the x (*interfacial all atom RMSD*) and y (*interface score*) cut off for eliminating the outliners. Among the successful simulations, the R1C and R3N PPIs have lower (more negative) *interface score* than the R2C and R3C PPIs, suggesting R1C and R3N PPIs are more favourable. The density contour plots revealed highly populated areas from each of the successful simulations. In N-SH2 PPIs (Figure 5.16 left), only R3N simulation was successful, and a single high density region was revealed from the analysis. In the C-SH2 PPIs (Figure 5.16 right), two high density regions were revealed from each of the C-SH2 analyses. Representative models identified from each of the high density regions are shown and discussed below.



Figure 5.16: Scatter and density contour plots from GAB1 586-592 to PI3K SH2 domain simulations. The scatter plots demonstrate the relationship between *interfacial all atom RMSD* and *interface score* for each of the simulation. Each dot in the scatter plot represents a model from the ensemble. A marginal distribution of the two variables is shown at the x and y axis. Density contour plots show 90% of the data presented in the corresponding scatter plot. Contour lines connect points of equal density in this graph. Of note, R1N and R2N simulation were unsuccessful.

Table 5.5: The values for 90% of the x and y distributions for each GAB1 586-592 to PI3K SH2 domain simulation. Models within these ranges were included in the density contour plot, models outside these values were treated as outliners and eliminated from the analysis. The *interfacial all atom RMSD* measures the geometric changes of the interfacial residues between starting and final structure. The *interface score* describes the free energy contributed by interface residues of each given complex, with negative values representing favourable contacts.

	Interfacial all atom RMSD (Å)	Interface score
GAB1 586-592 to R1N	Failed	Failed
GAB1 586-592 to R1C $$	x<4.5	y<-15
GAB1 586-592 to R2N $$	Failed	Failed
GAB1 586-592 to $R2C$	x<5.0	y<0
GAB1 586-592 to R3N $$	x<4.5	y<-15
GAB1 586-592 to R3C $$	x<4.5	y<0

The models identified from the high density regions from the PIK3R1 C-SH2 and PIK3R2 C-SH2 analyses are shown in Figure 5.17. In the R1C analysis, three representative models were identified, with one representing region A and two representing region B. Of the three structures, all have different peptide geometries and all highlighted the importance of 24L and 56V for stabilising the N-terminal peptide, the critical role of 58H is to secure the pY backbone orientation. In terms of geometrical differences, the pY side chain in structure A is facing in the SH2 domain (away from the plane), whereas the pY side chain in B1 and B2 are facing the C-terminal end of the peptide. This led to the formation of several favourable contacts between the phosphate group in structure A and loop5 residues (38R, 40S, 41S). In structure B1 and B2, as the pY side chains are not pointing toward loop5, the phosphate group in these two structures form one or none hydrogen bond with the residues from the loop region, resulting in a higher *interface score* (less stable complexation).

The two representative structure shown at the bottom of Figure 5.17 were identified from the high density regions from R2C PPI simulation. In structure A, the N-terminal peptide (pY-3 to pY) is parallel to the beta-C strand, with the peptide interacting with 43D, 44T, 45K, 46H and 48V from the beta-C strand. In structure B, the N-terminal peptide is pointing slightly towards the alpha-A helix, therefore, apart from binding to the residues from the beta-C strand, the peptide in structure B also interact with 8R and 12E from alpha-A helix.



Figure 5.17: Representative structures revealed from GAB1 586-592 to SH2 domains in PIK3R1 and PIK3R2. Two high density regions were identified from R1C simulation (top), in which one and two binding modes were revealed from region A and B, respectively. Two high density regions were revealed from R2C simulation (bottom) and each revealed a binding mode. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

The representative structures revealed from R3N (top) and R3C (bottom) docking are shown in Figure 5.18. The two structures, A1 and A2, revealed from R3N analysis have similar backbone orientations, with both of the N-terminal peptides interacting with residues from the beta-C strand and both of the C-terminal peptides interacting with residues from the alpha-C helix. Alpha-C helix is a short helix connected to the alpha-B helix by a short loop region, and the structures revealed from the analyses suggest that alpha-C helix is essential for binding to the proline at pY+2. Despite the pY side chain orientation in the two structures are different (one pointing toward the C-terminal peptide and one pointing toward the N-terminal peptide), neither of the phosphate group form any favourable contact with the SH2 domain, and therefore no differences in the *interface score* (free energy of binding) were observed between the two structures.

The bottom of Figure 5.18 shows the two representative structures for region A and B from R3C docking simulations. Different from the previous analyses, the two structures have quite distinct N-terminal orientation. In structure A, the backbone of glutamic acids and asparagine at the pY-3 and pY-1 interact with residues from the alpha-A helix and beta-C strand. Whereas the N-terminal residues in structure B are in a very different geometry, reducing the contacts with the residues in alpha-A helix and beta-C strand. In terms of the pY side chain in structure A and B, both fit into the positively charged pocket and bind to 26R and 28S at the surface of the SH2 domain. While the geometries of the two C-terminal peptides are different, in both structures the pY+2 proline sits in a shallow pocket, introducing a "bend" to the peptide backbone so that pY+1 and pY+3 residues can fit into a hydrophobic pocket, respectively.



Figure 5.18: Representative structures revealed from GAB1 586-592 to SH2 domains in PIK3R3. A high density region was identified from R3N simulation (top), revealing two distinct binding modes. Two high density regions were revealed from R3C simulation (bottom), with each revealing a binding mode. Peptides are shown in magenta sticks and SH2 domains are shown in green cartoon. N- and C-end of the peptides are labelled.

# 5.5 Molecular recognition of SH2 domains in PI3K

To summarise the molecular recognition features of SH2 domains in PI3K regulatory subunits, the four pY peptide sequences examined in this chapter were aligned and their orientations in relation to the SH2 domain were reviewed. The peptide sequence alignment (Figure 5.19A) shows the four pY peptides represent a conserved pattern, with residues N-terminal to pY generally being polar and negatively charged, and residues at pY+1 and pY+3 are hydrophobic and are separated by a proline or a hydrogen bond making residue. Given that most of the simulations presented in this chapter were successful and favourable (*interface score* < 0), the results suggest that SH2 domain in PI3K regulatory subunits prefer peptide sequences with a motif of negative-x-polar/negative-pY-hydrophobic-y-methionine (where x is any residues, y is a residue that can provide backbone conformational constraints). This observation extends the existing knowledge that SH2 domains in PI3K recognise a specific pY-X-X-M consensus sequence [411, 415, 416], and suggests that residues N-terminal to the pY (e.g. pY-1, pY-2, pY-3) also contribute to the binding.

Apart from the sequence preferences, the above analyses also demonstrated that SH2 domains have geometric preferences. In general, two types of binding geometries were revealed: parallel binding and orthogonal binding (Figure 5.19 B). In parallel binding, the N-terminal region of the peptides is in parallel to the beta-C strand and interacts with several residues from beta-C strand. In orthogonal binding, however, the peptide is positioned orthogonal to the beta-C strand and form hydrogen bonds with residues from the alpha-A helix. While orthogonal binding to the beta-C strand has been seen in many SH2-pY crystal structures [417, 418], including PIK3R1 SH2 complexes (e.g. PDB 5AUL), the results presented here suggested that this is not a universal mechanism and other binding geometries can also be achieved. Moreover, the above analyses also demonstrate two types of pY side chain geometries (Figure 5.19 C), with one pointing away from the plane and the other one pointing toward the C-terminal peptide. If the pY is pointing away from the paper plane, it generally fits into a positively charged pocket and forms electrostatic interactions (via its phosphate moiety) and van der Waals contacts (via its aromatic moiety) with the residues that made up the pocket.

Moreover, a common feature shared among most PI3K SH2 complexations is the hydrophobic contacts at the C-terminal regions. Specifically, C-terminal residue pY+1 and pY+3 fit into two deep hydrophobic pockets, forming hydrophobic interactions with SH2 domains (Figure 5.19 D).

It has been previously reported by Songyang *et al.* that SH2 domains in PIK3R1 have relatively conserved pY+3 pocket which strongly select methionine at +3 [415]. However, their results suggest that the pY+1 pocket in PIK3R1 is less selective and would interact with other nonmethionine hydrophobic residues. Their observation is in agreement with the results presented in this chapter, where both methionine and valine could fit into the pY+1 pocket, and only methionine is demonstrated to fit into the pY+3 pocket. Moreover, the residue between pY+1and pY+3 also present an important position for the complexation. This pY+2 position is thought to be providing a bend, or some degree of conformational constraint to reduce the flexibility of the backbone of C-terminal regions. In the models shown above, this was provided either by proline or a hydrogen making residue (e.g. aspartic acid).



Figure 5.19: Characterisation of pY peptide binding modes. (A) Sequence alignment of the four peptides considered in this chapter. Amino acids are coloured according to their charge, polarity, and hydrophobicity. (B) Two peptide binding geometries were observed from the analyses above. In a given complex, a peptide can bind either parallel or orthogonal to the beta-C strand. (C) Two distinct pY orientations were also observed, with one pointing toward the SH2 domain and one pointing towards the C-terminal peptide. (D) In most PI3K SH2 domain complexes, C-terminal pY+1 and pY+3 residues fit into a hydrophobic pocket each.

The above analyses have also highlighted the importance of beta-C strand, especially the 4th residue, in peptide binding. As mentioned before, the pY peptides examined in this chapter all have a negatively charged N-terminal region, and the residues from these regions often interact with the charged or polar residues from the beta-C strand. Importantly, the carboxyl group from pY-1 and the amino group from pY+1 in every representative structure interact with the 4th residue from the beta-C strand, suggesting that the 4th residue in the beta-C strand is vital for stabilising the pY backbone (Figure 5.20). The 4th residue of the beta-C strand in N-SH2 is leucine, and its backbone forms two strong (less than 2.5 Å) hydrogen bonds with pY-1 and pY+1 in all representative models. On the other hand, the 4th residue of the beta-C strand in the C-SH2 domain is histidine, which also binds to the carboxyl group from pY-1 and the amino group from pY+1. The aromatic side chain of the histidine in C-SH2 also provides a potential site for  $\pi$ - $\pi$  stacking interaction with the aromatic ring in the pY side chain.



Figure 5.20: Beta-C strands in SH2 domains in PI3K regulatory subunits. The residues that make up beta-C strand in N-SH2 (top left) and C-SH2 (top right) domains are aligned. The beta-C strand in R2C and R3C are four residues shorter than the beta-C strands in other SH2 domains. The fourth residue in the N-SH2 domain and C-SH2 domain are conserved, forming a stabilising backbone-backbone contacts with pY-1 and pY+1 positions (bottom).
# 5.6 Discussion

In this chapter, the putative peptide sequences from KIT and GAB1 were docked onto the SH2 domains in PI3K regulatory subunits. The results suggest that most sequences identified from the previous chapter can potentially interact with both N-SH2 and C-SH2 domains in the PI3K regulatory subunits. In the following sections, the biological context of GAB1-PI3K and KIT-PI3K interactions are discussed as well as the limitations and possible extensions of this work.

#### 5.6.1 Biological roles

This chapter looked to characterise GAB1-PI3K and KIT-PI3K interactions. While the exact biological roles of these PPIs in BLBC are unclear. It has been shown by previous studies that in normal physiological condition, tyrosine phosphorylated GAB1 interacts with SH2 domain in PIK3R1, and the interaction leads to plasma membrane recruitment and activation of PI3K. Activated PI3K then catalyses the production of phosphatidylinositol-3,4,5-triphosphate (PIP3) by phosphorylating phosphatidylinositol-4, 5-bisphosphate (PIP2). PIP3 binds and activates PH-containing enzymes (i.e. pyruvate dehydrogenase kinases), and consequently activate AKT. Previous studies showed that GAB1-PIK3R1 interaction is essential for mediating the PI3K/Akt signalling pathway and mutating the PI3K binding site of mammalian GAB1 can lead to incapacitation of PI3K/Akt signalling pathway and several other signalling systems [419–422].

The PI3K/Akt signalling pathway regulates several essential cellular processes including metabolism, proliferation, growth, and survival. It is one of the most frequently dysregulated pathways in cancers and have been shown to promote tumorigenesis and tumour development [423, 424]. In intrahepatic cholangiocarcinoma, Sang *et al.* showed that expression of GAB1 is positively correlated with lymph node metastasis and TNM stage and interfering GAB1 leads to a reduction of PI3K/Akt activity and a decrease in cell proliferation [425]. It has also been shown that GAB1-mediating PI3K/Akt pathway regulate cancer proliferation and metastasis in thyroid tumours [426]. In breast cancer, while GAB1 upregulation has been linked to breast cancer metastasis, results from Wang *et al.* suggested that the metastasis is independent from the PI3K/Akt pathway and is regulated by other GAB1-mediating signalling pathways [328]. These findings suggest that GAB1-mediated PI3K activation and its downstream pathways may

have different biological roles in different cell types and cancers, and therefore deregulation may result in different consequences. Understanding the specific involvement and biological role of GAB1-PI3K.

PI3K can also be activated by receptor tyrosine kinases (RTKs), including KIT. The KIT-PIK3R1 interactions explored in this chapter are involved in this process. KIT is a transmembrane receptor that binds to stem cell factors (SCFs). Once activated, KIT undergoes dimerization and auto phosphorylation on multiple tyrosine residues. The phosphorylated tyrosine 721 in KIT has been previously shown to bind to the PI3K regulatory subunit 1 (PIK3R1) and regulate downstream signalling pathways that are responsible for controlling cell adhesion, survival, proliferation [427–429]. Previous study demonstrated that mutating Y721 in KIT does not rescue cells from Bad-induced apoptosis in U2-OS cells (osteosarcoma cells) [427]. In bone marrow mast cells (BMMCs), mutation at Y719 in KIT reduces the SCF-induced proliferation and protection from apoptosis. A reduction in Akt activation was also observed [430]. It has also been shown that disrupting the KIT-PI3K interactions by mutating Y721 in KIT resulted in a significant reduction in growth and tumorigenicity of cells expressing the KIT kinase catalytic domain mutant [431–433]. Results from these studies suggested that while KIT-PI3K interactions might not be required, they indeed play a role in regulating SCF-mediated signalling pathways and cellular responses (i.e. proliferation, survival). It is hypothesised that interfering KIT-PI3K interactions does not only affect PI3K kinases, but also blocked recruitment of other proteins that bind to KIT via PIK3R1 [434]. This suggests that KIT-PI3K interaction can trigger multiple signalling pathways and regulate different biological processes, highlighting the complexity of KIT-PI3K mediating processes. Understanding the involvement and biological roles of KIT-PI3K interaction in oncogenic signalling is an ongoing challenge.

#### 5.6.2 Role of water in protein interactions

The docking simulations presented in this chapter were performed *in vacuo*, with all the explicit water molecules being excluded. While removing water molecules provides several advantages such as less computational time and preventing distortion during the docking procedure, it makes it challenging to determine where interfacial waters might be essential in a given complexation. In a study done by Ahmed *et al.*, they showed that 21% of the water molecules in X-ray crystal structures have a role of bridging the interactions between two proteins and stabilising the PPIs [435]. Moreover, Roberts and Mancera showed that the inclusion of critical

water molecules could significantly improve the accuracy of docking predictions [436]. While the above studies are likely to be system-specific, it is still important to note that the removal of water molecules might have a biased contribution to these docking results.

To investigate the involvement of the water molecules in PI3K SH2 domain binding, future work could perform equivalent simulations with the inclusion of surface-bound and conserved water molecules from the relevant crystal structures. A comparison between *in vacuo* and water-bridging systems could reveal the importance of hydration shell around the protein binding site and has these water molecules contribute to stabilising or destabilising the complexation [437]. As there is a negatively charged phosphate group in the system, it is expected that water molecules would desolvate and play a role in stabilising the exposed charged groups at the protein surface [438]. It is also expected that the inclusion of surface-bound water molecules would influence the both kinetics and the thermodynamics of the protein complexation [439], and further studies could use molecular dynamics tools to explore the hydration at the SH2 domain interface.

#### 5.6.3 Exploring flanking regions

GAB1-PI3K and KIT-PI3K interactions explored in this chapter are thought to be mediated by SH2 domains. The SH2 domain is a small binding domain that modulates many signalling PPIs [403, 409]. An early work by Songyang et al. used a peptide library to predict the preferential binding motifs for 25 SH2 domains [415, 440]. Their results demonstrated that different SH2 domains recognise different 4-mer peptide motifs (pY to pY+3). In particular, they showed that SH2 domains from PI3K bind preferentially to  $pY\phi X\phi$  motif (where X is any residues and  $\phi$  is a residue with a hydrophobic side chain). In addition, recent research has also demonstrated the importance of the motif flanking regions in modulating the SLiM-mediated interactions [441]. This was in agreement with the docking results presented in this chapter, where motif flanking residues, in particular the residues N-terminal to the pY (pY-1 to pY-3), form several hydrogen bonds with residues from the beta-C strand, potentially modulating the complexation. This should inform and motivate future work in exploring the different lengths of flanking regions, both C-terminal and N-terminal from the pY motif. Further work into the pY flanking regions should provide insight into binding selectivity among the three PI3K regulatory subunits. It should be mentioned that, different peptide docking protocols are designed to handle different peptide length, and using a peptide that is out of the desired range might lead to uninformative

results [442]. For example, Rosetta FlexPepDock was deigned to handle peptides with 5-12 residues, and peptide longer than 12 residues might hinder site detection, and indeed peptides shorter than five residues might give rise to weak and noisy results.

#### 5.6.4 Future challenges and perspectives

While the work presented in this chapter provides structural insight into the molecular recognition of the KIT-PI3K and GAB1-PI3K PPIs, the high sequence similarity and structural conversation between SH2 domains in PI3K regulatory subunits pose a significant challenge for the development of isoform-selective inhibitors. With an increasing appreciation about the complex signalling events regulated by PI3Ks, the next challenge would be to uncover the divergent roles of each PI3K regulatory isoforms in different signalling contexts. Moreover, as an extension of this work, future work should also explore the potential pharmacophore of each PI3K regulatory subunits. Studies such as multiple pharmacophore mapping would reveal isomer-selectivity determinants and consequently, inform the development of selective and potent inhibitors. It is important to note that other factors, such as protein localisation are also likely to play an essential role in determining specificity during the formation of SH2-mediating PI3K complexes. In the next chapter, the cellular protein localisation in Basal-like breast cancer cells will be explored using cellular assays.

## 5.7 Key findings

- Previous reported pY peptide sequences from KIT and GAB1 binds to both N- and C-SH2 domain in PIK3R1, PIK3R2 and PIK3R3, with no clear preferences.
- pY peptide can bind parallel or orthogonal to the beta-C strand in the PI3K SH2 domains.
- SH2 domains in PI3K has a preferred motif pattern of negative-x-polar/negative-pYhydrophobic-y-methionine (x is any residues; y is a residue that can provide backbone conformational constraints).

# Chapter 6

# **Experimental Validation**

# 6.1 Introduction

While the previous chapter characterised the molecular recognition of SH2-peptide binding for KIT-PI3Ks and GAB1-PI3Ks PPIs, this chapter will explore endogenous KIT and PI3K regulatory subunits in four different breast cell lines and examine whether KIT colocalises with PIK3R1 and PIK3R2 in Basal-like breast cancer.

#### 6.1.1 Protein kinases in cancer

Human protein kinases are a large family of enzymes that can transfer a phosphate group from ATP to specific residues (typically hydroxylated) in their substrate proteins. Phosphorylation is a critical post-translational modification that introduces a negatively charged phosphate group to the surface of the substrate protein, changing the electrostatic distribution and potentially inducing conformational changes. The conformational changes can disrupt or encourage proteinprotein interactions, subsequently leading to a change of protein activity, cellular location, or association with other proteins [443]. Depending on the kinase cellular location, they can be classified into transmembrane receptor kinases and non-receptor kinases. Transmembrane receptor kinases consist of an extracellular ligand-binding domain, a transmembrane domain and an intracellular catalytic domain, whereas non-receptor tyrosine kinases locate to the cytosol, nucleus, or intracellular side of the cell membrane and contain a regulatory domain and a catalytic domain [444]. Transmembrane receptor kinases can be activated by extracellular ligands or a stimulus, and activation often leads to intracellular signalling cascades, including non-receptor kinases activation. This coordinated action of kinases and their substrate proteins allows transduction of extracellular and intracellular signals through the cytoplasm and towards the nucleus.

Protein kinases are only coded for by about 1.7% of the all human genome, however, they can phosphorylate more than 30% of cellular proteins [445], highlighting their importance in regulating and maintaining essential cellular processes such as proliferation, metabolism and apoptosis. However, deregulation (e.g. mutation, misregulation or chromosomal translocation) of kinases activities does lead to alterations in these processes and can contribute to oncogenesis, allowing abnormal tumour cell proliferation, differentiation, and apoptosis inhibition [446–448]. Deregulated protein kinase activate pathways have become attractive drug targets for cancer therapies. In the past 20 years, several kinases (e.g. breakpoint cluster region and tyrosine-protein kinase ABL1 fusion protein (BCR-ABL), EGFR, ERBB2 (HER2), KIT, VEGFR, anaplastic lymphoma kinase (ALK), serine/threonine-protein kinase B-Raf (BRAF)) have been demonstrated to be proven therapeutic targets to treat cancers [449, 450].

Deregulation of protein kinases has also been identified in different breast cancer subtypes, and some have led to successful developments of targeted treatment. An example is HER2-positive breast cancer, where HER2, a transmembrane receptor tyrosine protein kinase, is constitutively activated [451, 452]. Over-expression of Her2 has been linked to dysregulation of cell proliferation in HER2-positive breast cancer tumours. Several targeted therapeutic molecules (e.g. lapatinib, trastuzumab and pertuzumab) have been developed to target Her2, with strategies demonstrating significant efficacy in treating Her2-positive patients [453].

While some aberrant protein kinase activities (e.g. PTEN [454], MAPK[455], aurora kinase (Auroa)A/B [456, 457]) have been identified in Basal-like breast cancer, none of these discoveries has led to promising clinical trials, highlighting the ongoing challenge of identifying a potent molecular target (or combinations of targets) for treating Basal-like breast cancer. Previous chapters have demonstrated the importance of KIT-PI3K regulatory PPIs in mediating the LCK and STAT1 associations. To explore whether KIT-PI3K regulatory subunits are crucial to Basal-like, this chapter describes work that aimed to reveal the cellular characteristic of KIT, a transmembrane tyrosine kinase, in Basal-like cells and its association with its binding partners (PIK3R1 and PIK3R2) in the hope of shedding more light on the complexity of kinase signalling network underlying Basal-like subtype.

#### 6.1.2 KIT

Transmembrane receptor tyrosine kinases (RTKs) are a family of kinases that are able to phosphorylate tyrosine residues in their target proteins in a ligand dependant manner. According to their structures and binding ligands, RTKs can be classified into 20 classes and KIT, is a member of class III transmembrane receptor tyrosine kinase. It has been demonstrated that KIT and other members of class III RTKs, platelet-derived growth factor receptor (PDGF) proteins, are highly expressed in breast cancer [458]. Moreover, work done by Roswall *et al.* have suggested that members of class III RTKs are involved in key signalling pathways that evolve breast tumours towards a more Basal-like phenotype [459]. In a separate work, they have also demonstrated that high expression of type III RTKs is associated with poor prognosis in ER-negative breast cancer [460]. KIT alone has also been found to be associated with poor prognosis in Basal-like, demonstrated by several studies [313, 315, 316, 461]. While these previous studies have highlighted and demonstrated the involvement of type III RTKs, including KIT, in breast cancer oncogenesis, their role and how their activities regulate oncogenic pathways in breast cancer especially Basal-like remains unclear.

#### 6.1.3 PI3K regulatory subunits

Phosphoinositide 3-kinases (PI3Ks) is a family of enzymes that are able to phosphorylate phosphatidylinositol 3, 4-bis-phosphate (PIP2) lipids to phosphatidylinositol (3,4,5)-trisphosphate (PIP3). Depending on their structures and functions, the PI3K family is divided into four classes, with Class I being the most well-studied class. Class I PI3Ks are heterodimeric containing a regulatory and a catalytic subunit and based on their sequence similarity, Class I PI3K can be further divided into IA and IB. Class IA PI3Ks is a heterodimer containing a p110 catalytic subunit and a p85 regulatory subunit, and the three PI3K regulatory subunits (PIK3R1, PIK3R2 and PIK3R3) considered in this thesis belongs to this class. Class IA PI3K can be activated either by G-protein-coupled receptors (GPCRs) or receptor tyrosine kinases (RTKs) (Figure 6.1). Upon activation, PI3K phosphorylates PIP2 and trigger a series of cell growth and cell survival pathways, such as the mTOR signalling pathway and Akt signalling pathway [462].



Figure 6.1: Schematic illustration of Class IA PI3K activation. Class IA PI3Ks consist of a p110 catalytic subunit (e.g. PIK3CA) and a p85 regulatory subunit (e.g. PIK3R1). Once its extracellular ligand activates receptor tyrosine kinase (e.g. KIT) or G-protein-coupled receptor (GPCR), the p85-p110 complex is recruited to the membrane via different mechanisms (either mediated by tyrosine phosphorylation of RTKs or interaction with G-proteins  $\beta/\gamma$ ). Once the p85 regulatory subunit is bound to KIT or GPCR, it relaxes its inhibitory effect on the p110 catalytic subunits and activates class IA PI3K. Once activated, PI3K can phosphorylate PIP2 into PIP3. PIP3 then binds to its downstream proteins that are involved in cell growth and cell survival pathways such as AKT. The involvement of PIK3R2 and PIK3R3 remains unclear.

The p85 regulatory subunits play an essential role in mediating PI3K activation, and mutation or alteration of p85 has been associated with cancer progression. For example, overexpression of PIK3R2 in breast cancer have been demonstrated and linked to oncogenic transformation of primary fibroblasts [463, 464]. In contrast to PIK3R2, findings from Thorpe *et al.* suggest that PIK3R1 play a tumour-suppressive role in oncogenic transformation. They demonstrated that the loss of PIK3R1 in breast cancer increases the recruitment of p85-p110 heterodimers to RTKs, driving tumour progression and oncogenic transformation [465]. This highlights that, although PI3K regulatory subunits have a similar structural arrangement and high sequence similarity, they can have distinct roles and give rise to diverse cellular responses. To gain more insight into the roles of different variant of PI3K regulatory subunits in BLBC, this work will explore PIK3R1 and PIK3R2's expression pattern, subcellular localisation and their association with KIT and phospho-KIT in four different breast cell lines. PIK3R3 was not included in this research due to the lack of commercial antibody.

#### 6.1.4 Aims

This chapter aimed to explore endogenous KIT and PI3K regulatory subunits and their associations in four breast cell lines. This includes:

- revealing their expression patterns and identifying their subcellular localisation;
- determining whether KIT and PI3K regulatory subunits colocalised in Basal-like cell lines;
- exploring the effect of KIT inhibitors in Basal-like.

#### 6.1.5 Chapter overview

The overview and summary of the steps and experiments carried out in this chapter are shown in Figure 6.2.



Figure 6.2: Chapter overview.

The details of the experimental protocols and reagents used to perform immunofluorescence (IF) and proximity ligation assay (PLA) are described in Section 6.2. The KIT/pKIT and PIK3R1/2 expression pattern were revealed using IF, and their colocalisation were detected using ! (!). Results from IF and PLA analyses are shown in Section 6.3. Results from IF experiments are shown in Section 6.3.1. In this section, protein expression pattern and their sub-cellular localisation are discussed. Results from PLA experiments are demonstrated in Section 6.3.2, where protein colocalisation are explored and compared between cell lines and different protein

regions (epitopes). Finally, KIT inhibitors and their potential effect on BLBC are discussed speculatively in Section 6.4.

# 6.2 Materials and methods

The endogenous KIT and PI3K regulatory subunits were studied using cell-based assays in this chapter. Protocols and steps regarding cell culture and cell maintenance are described in Section 6.2.1 to 6.2.5. Indirect dual immunofluorescence (IF) assay was used to reveal protein localisation and the details of the IF protocol are described in section Section 6.2.6. Colocalisation of KIT/pKIT and PIK3R1/2 were explored using *in situ* proximity ligation assay (PLA), and the experimental procedure of the experiments regarding PLA is described in Section 6.2.7. Finally, the choice of antibody combinations used for colocalisation experiments is explained and described in Section 6.2.8.

#### 6.2.1 Cell culture

Four cell lines were cultured in this study, including three triple negative breast cancer (TNBC) cells and a healthy breast epithelial cell line. Cell line MDA-MB-231 and HBL100 were gifted by Dr Josephina Sampson (University of Leeds and University of Leicester). Cell line MDA-MB-468 and LGI1T were gifted by Prof Valerie Speirs (University of Leeds and University of Aberdeen). MDA-MB-231, MDA-MB-468 and HBL100 are commercial cell lines, and have all been previously registered. Their short tandem repeat (STR) profiles and descriptions can be found on BioSample database (https://www.ncbi.nlm.nih.gov/biosample) with their corresponding BioSample IDs: MDA-MB-231 (BioSample: SAMN03472205), MDA-MB-468 (BioSample: SAMN03473056), HBL100 (BioSample: SAMN03472914). Cell line LGI1T was isolated by Prof Valerie Speirs' group, from a recurrent breast tumour from an elderly patient who received tamoxifen as primary therapy. The primary tumour was classified as a lymph node negative,  $ER\alpha$  negative and PR negative squamous cell carcinoma. Phenotypic characterisation by immunofluorescence (IF) revealed expression of cytokeratins CK5/6 and CK14, but not ER $\alpha$  or HER2 or the luminal markers CK18 and CK19, which is consistent with Basal-like subtype. LGI1T cell line represents Basal-like phenotype. LGI1T has not yet been verified at the time of writing this thesis. None of the cell lines used in this project are known to be misidentified (https://iclac.org/databases/cross-contaminations/). Please note that the cell lines were not

authenticated in this project, but it is a good practice to authenticate cell lines prior any experiments to ensure the cells are not misidentified, cross contaminated, or genetically different from the original cell stock. Cell lines can be authenticated using short tandem repeat (STR) profiling method, which is a widely used and accepted authentication method.

Details of these cell lines are summarised in Table 6.1. All cell lines were maintained in a humidified incubator at  $37 \,^{\circ}$ C with 5% CO<sub>2</sub> in 96mm tissue culture dishes from TPP. Cell lines were tested regularly (monthly) for mycoplasma infection using a highly sensitive and specific PCR-based assay (EZ-PCR Mycoplasma kit, Geneflow).

Table 6.1: Details of the four cell lines used in this chapter. Classification adapted from [48, 217, 466, 467]. Culture media culture media dulbecco?s modified eagle medium (DMEM) and **rpm1 1640!** (**rpm1 1640!**) were supplied by Gibco. Keratinocyte medium Kit (product name: Keratinocyte-SFM Medium (Kit) with L-glutamine, epidermal growth factor (EGF), and bovine pituitary extract (BPE), catalog number: 10144892) comes with pre-measured BPE (bovine pituitary extract) and EGF (epidermal growth factor) and was supplied by Gibco. Fetal bovine serum (FBS) was also supplied by Gibco.

Cell line	Classification	Immuno-profile	Growth Medium	Weekly Passage
MDA-MB-231	Claudin-Low	ER-, PR-, HER2-	$\begin{array}{l} \text{DMEM (glutamax)} \\ + 10\% \text{ FBS} \end{array}$	1:12
MDA-MB-468	Basal-like	ER-, PR-, HER2-	$\begin{array}{l} \text{RPMI 1640} \\ + 5\% \text{ FBS} \end{array}$	1:10
LGI1T	Basal-like	ER-, PR-, HER2-	$\begin{array}{l} \text{Keratinocyte} \\ + \text{ BPE} + \text{EGF} \end{array}$	1:10
HBL100	Healthy (breast epithelial)	N/A	$\begin{array}{l} \text{DMEM (glutamax)} \\ + 10\% \text{ FBS} \end{array}$	1:8

#### 6.2.2 Cell passage

Cells were initially washed with 1 x phosphate-buffered saline (PBS) then incubated for 5 minutes in 2ml of Trypsin. Cells were resuspended in their appropriate growth media and dilute for passage (see weekly passage in Table 6.1. LGI1T cells were resuspended in RPMI 1640 media containing 5% FBS to deactivate the Trypsin and then centrifuged for 5 minutes at 1100 rpm using Eppendorf 5804 R Centrifuges with A-4-44 Model Rotor, after which the cell pellet was resuspended in Keratinocyte growth media and diluted in the appropriated amount of media for passage.

#### 6.2.3 Cell freezing

At least eight stock vials of frozen cells were kept for each cell line to allow return to a standard passage number if and when required. All cells were frozen in freezing media (5% dimethyl sulfoxide (DMSO) + fetal bovine serum (FBS)). Cells at 80% confluency were washed in 1 x PBS and trypsinised as previously described in Section 6.2.1. Cells were resuspended in 10ml of media containing FBS and centrifuged at 1100 rpm for 5 minutes to pellet. The cell pellet was collected and resuspended in 1 ml of freezing media, which then was transferred into each cryovial. These cryovials were then moved to an isopropanol filled container to ensure optimal cooling rate at -80 °C overnight. The following day frozen stocks were transferred to liquid nitrogen for long term storage until required.

#### 6.2.4 Thawing cells

Cultured cells were discarded at a maximum of 25 passages and replaced with a new stock vial. When cells stocks were thawed from liquid nitrogen, they were cultured for at least one passage before use in experiments. Frozen vials were thawed quickly by placing in 37 °C water bath for 30 seconds. Cells were mixed with 9 ml of their respective media and centrifuge at 1100 rpm for 5 minutes to pellet. The media supernatant was removed, and the cell pellet was resuspended in 10 ml of its respective media. The cell suspension was placed in a 96mm tissue culture dish in an incubator 37 °C with 5% CO<sub>2</sub>.

#### 6.2.5 Cell seeding and fixing

Sterilised acid-etched glass coverslips were placed into the wells of a sterilised 12 well plate. Cells were plated onto the coverslips with  $0.2 \ge 10^6$  cells per coverslip. Cells were allowed to fully adhere for 24 hours. All cells were washed twice in 1 x PBS before fixing. HBL100 and MDA-MB-468 were fixed with paraformaldehyde which requires incubating in 3.7% paraformaldehyde for 10 minutes under the fume hood. MDA-MB-231 and LGI1T were fixed with iced cold methanol. Paraformaldehyde fixed cells were stores in 1x PBS at 4 °C, and methanol fixed cells were stored in methanol at -20 °C.

#### 6.2.6 Indirect dual immunofluorescence protocol

The indirect dual immunofluorescence (IF) method was used to examine protein localisation. Details of the antibodies used, their working concentrations and diluents used are detailed in Table 6.2. All washes were performed using PBS. MDA-MB-468 and HBL100 were fixed with paraformaldehyde and permeabilised with 5% triton. MDA-MB-231 and LGI1T and were fixed with iced cold methanol. Fixed and permeabilised cells were washed three times before incubating in blocking solution (3% bovine serum albumin (BSA) in PBS) for 1 hour. The blocking solution was removed, coverslips were washed twice with PBS. The coverslips were then incubated in  $40\mu$ l of diluted primary antibody at 4 °C overnight. The next day, primary antibodies were removed and washed twice in 1xPBS.  $40\mu$ l of secondary antibodies (1:200 in 1% BAS) was added onto the coverslips and incubated for an hour in the dark. Secondary antibodies were Alexa Fluor-488 and -594 goat anti-rabbit and goat anti-mouse IgGs (Invitrogen). After secondary incubation, coverslips were washed three times and mounted onto labelled slides with cell side down using Duolink<sup>™</sup>In Situ Mounting Medium with 4, 6-diamidino-2-phenylindole (DAPI). Slides were then stored at 4 °C in the dark until image acquisition. Cells were visualised on a Zeiss LSM880 + Airyscan Inverted confocal microscope using a 40x oil objective (numerical aperture, 1.4). Z-stacks comprising of 10-20 x 0.3  $\mu$ m sections were acquired. Images were analysed using Fiji (v.2.0.0-rc-69/1.52p). Images were acquired within seven days of staining. Coverslips were stored in the dark at 4 °C until image acquisition. A negative and a positive control were included in all immunofluorescence experiments. AuroaA and TPX2 microtubule nucleation factor (TPX2) were used as a positive control where clear staining of each protein and colocalisation of both proteins were observed in mitosis cells. In negative slides, the primary antibody was omitted, and only DAPI staining was observed, and no staining was observed for the protein of interest. Images captured and presented from immunofluorescent experiments were representative of the whole slide. Three biological replicates were performed for every experiment, and results shown are a typical subset.

Antibody details (Company,CatID)	Species	<b>Epitope</b> ProteinID_sequence (domain/region)	Assay	Working concentration
KIT-Y703 (CellSignalling, D13A2)	Rabbit	P10721_Y703	IF/PLA	1:250
KIT-pY721 (ThermoFisher, 44-494G)	Rabbit	P10721_pY721	IF/PLA	1:250
PIK3R1-SH2 (abcam, ab189403)	Mouse	P27986_AA159-388 (N-SH2)	IF/PLA	1:100
KIT-cytoplasmic (CellSignaling, 3308S)	Mouse	P10721 (Cytoplasmic tail)	IF/PLA	1:800
PIK3R1-SH3 (ThermoFisher, PA5-32550)	Rabbit	P27986 (SH3)	IF/PLA	1:100
PIK3R2- Rho-GAP (abcam, ab180967)	Rabbit	O00459_100-200 (Rho-GAP)	IF/PLA	1:500

Table 6.2: **Details of antibodies used in this chapter.** All antibody was diluted in 3% BSA. **IF**: Immunofluorescence; **PLA**: proximity ligation Assay.

While dual immunofluorescence staining could also be used to identify colocalisation between two proteins, the protein staining presented in this work were too weak and/or too noisy to perform meaningful colocalisation analysis. Therefore, protein colocalisation between proteins is examined by another assay, Proximity Ligation Assay (PLA).

#### 6.2.7 In situ Proximity Ligation Assay protocol

Proximity Ligation Assay (PLA) was used to detect protein colocalisation. Cells were fixed and permeabilised, as described in Section 6.2.4. The same four pairs of antibodies from the immunofluorescence experiments were tested. Cells were washed with 1xPBS twice before they were incubated in 3% BSA blocking solution for 30 minutes to an hour. Diluted primary antibodies raised in two different species were then added on to the slides and incubated overnight at 4 °C. The next day, primary antibodies were tapped off, and the coverslips were washed twice with 1xPBS. PLA probes MINUS and PLUS stock (Duolink®, Sigma) were then diluted 1:5 in 3% BSA and left at room temperature for at least 20 minutes for activation. Diluted PLA probe solution was then added to the coverslips for an hour incubation in a humidity chamber at 37 °C.Once the PLA probe incubation is completed, the PLA probe solution was removed, and the slides were washed in Buffer A (Duolink, Sigma) twice. To make up the ligation solution, ligation stock (Duolink, Sigma) was diluted 1:5 in MilliQ water. During the washes, ligase (Duolink, Sigma) was retrieved from freezing using a freezing block. Ligase was then carefully added to the 1 x ligation buffer at 1:40 dilution and mixed well. After the washes, excess wash buffer was removed, and the ligation solution was applied to all samples (40  $\mu$ l/slide). The slides were incubated in a humidity camper for 30 minutes at 37 °C. Once the ligation incubation is completed, the remaining ligation solution is removed, and slides were washed in Buffer A (Duolink, Sigma) twice. To make up the amplification, amplification stock was diluted 1:5 in MiliQ water. During the washes, polymerase (Duolink, Sigma) was retrieved from freezing using a freezing block. Polymerase was then carefully added to the 1 x amplification buffer at 1:80 dilution and mix. After the washes, excess wash buffer was removed, and the amplification solution was applied to all samples (40  $\mu$ l/slide). The slides were incubated in a humidity camper for 100 minutes at 37 °C. Once the amplification incubation is completed, remaining amplification solution was removed, and the slides were then washed with 1x Buffer B (Duolink, Sigma) twice for 10 minutes and then washed with 0.01 x buffer B once for one minute. The slides were then mounted onto labelled slides with cell side down using Duolink<sup>™</sup> in situ Mounting Medium with DAPI. Slides were then stored at 4 °C in the dark until image capture. Images were acquired within ten days of staining.

Imaging was performed on a Zeiss LSM880 + Airyscan Inverted confocal microscope using a 40x oil objective (numerical aperture, 1.4). Z-stacks comprising of 10-20 x 0.3  $\mu$ m sections were acquired. Images were analyzed using Fiji (v.2.0.0-rc-69/1.52p). A negative and a positive control were included in all PLA experiments. AuroaA and TPX2 were used as a positive control. In negative slides, only a single primary antibody was added onto the coverslips. Images captured were representative of the whole slide. Three biological replicates were performed for every experiment and results shown are a typical subset. The number of PLA dots detected per cell was counted manually and data represent the mean of three independent experiments  $\pm$  SD. Statistical analyses were performed using one-way ANOVA analysis from Prism 7.0 software with the following levels of uncertainty: \*\*\*\*p<0.0001, \*\*p<0.001, \*\*p<0.01, \*\*p<0.05.

#### 6.2.8 Antibody combinations

In this work, four pairs of antibodies were examined (Figure 6.3). The first pair consists of PIK3R1-SH2 and KIT-Y703. The KIT-Y703 antibody (CellSignalling, D13A2) is a rabbit anti-

body that was produced to recognise the residues surrounding Tyr703 of total KIT protein. The PIK3R1-SH2 antibody (abcam, ab189403) is a mouse monoclonal antibody that is expected to bind to the N-SH2 domain in PIK3R1. Both antibodies in the first pair were raised to recognise the putative binding site or region close to the putative binding site. The second pair consists of PIK3R1-SH2 and KIT-pY721. The KIT-pY721 antibody was a rabbit polyclonal antibody (ThermoFisher, 44-494G) that was raised to recognise the phosphorylated tyrosine 721 in KIT. As discussed in the previous chapter, pY 721 is thought to be essential for the PIK3R1-KIT PPI. Different from the first two pairs of antibodies, the third pair of antibody consists of two antibodies (PIK3R1-SH3 and KIT-cytoplasmic) that recognise regions away from the putative binding sites, such as SH3 domain in PIK3R1 and the cytoplasmic tail in KIT. The PIK3R1-SH3 antibody (ThermoFisher, PA5-32550) is a rabbit polyclonal antibody that was raised by immunizing rabbits with a synthetic peptide derived from N-terminus of human PI3K p85. The KIT-cytoplasmic antibody (CellSignaling, 3308S) is a mouse monoclonal antibody that was raised by immunizing mouses with recombinant proteins containing the cytoplasmic domain of human KIT protein. Similar to the third antibody pair, the fourth antibody pair also contains two antibodies (PIK3R2-RhoGAP and KIT-cytoplasmic) that recognise regions away from the binding sites. The PIK3R2-RhoGAP antibody (abcam, ab180967) is a rabbit recombinant monoclonal antibody that recognises the Rho-GAP domain (residue 100-200) of PIK3R2. In this work, indirect dual immunofluorescence assay was used to detect the localisation of each protein and the colocalisation of these four pairs of antibodies were examined using a PLA assay. In the following sections, the first and second pair of antibodies that recognise the putative binding sites are referred to as bs-PIK3R1+KIT and bs-PIK3R1+pKIT, respectively. Similarly, the third and fourth antibody pairs that are expected not to interact with the putative binding sites are referred to as nbs-PIK3R1+KIT and nbs-PIK3R2+KIT, respectively.



Figure 6.3: Schematic illusion of antibody pairs examined in this chapter. Protein architectures were obtained from SMART (http://smart.embl-heidelberg.de). Antibodies are shown as red Y-shapes. Some antibodies were raised to recognise specific epitopes (monoclonal), and some were more generic and can recognise a large region of the protein (polyclonal). For example, antibody KIT-Y703 was raised to only recognise a single residues tyrosie703 in KIT, whereas antibody KIT-cytoplasmic was raised to bind to any cytoplasmic regions in KIT.

# 6.3 Results

#### 6.3.1 Protein sub-cellular localisation

The indirect dual immunofluorescence was used to examine the protein localisation. Expression of KIT-Y703, KIT-pY721, KIT-cytoplasmic, PIK3R1-SH2, PIK3R1-SH3 and PIK3R2-RhoGAP were evaluated in four different cell lines, including MDA-MB231, MDA-MB-468, LGI1T and HBL100. This experiment aimed to explore the localisation of the proteins in different breast cell lines.

The expression patterns of the proteins were first examined in MDA-MB-231, a Claudin-low cell line (Figure 6.4). Claudin-low is a small subset of triple-negative breast cancer and, compared to Basal-like, it is slow-cycling with lower proliferation levels [48, 468]. MDA-MB-231 is a highly aggressive, invasive and poorly differentiated cell line with TNBC characteristics (ER-, PR- and Her2-) [469, 470]. Different from Basal-like cell lines, MDA-MB-231 exhibits a low expression of Ki-67 and down-regulation of claudin-3 and claudin-4 [466]. Previous studies have shown that PIK3R1 is localised the cytoplasm in other cell lines (A-431 (human squamous carcinoma), U-2OS (human bone osteosarcoma epithelial cells), U-251MG (human malignant glioblastoma multiforme), https://www.proteinatlas.org/ENSG00000105647-PIK3R2/cell). PIK3R1 has also been seen in plasma membrane, cytosol, nucleus, endoplasmic reticulum and golgi apparatus (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PIK3R1&keywords=PIK3R1#localization). In MDA-MB-231, PIK3R1-SH2 and PIK3R1-SH3 expression were localised to the cytoplasm, with some strong nuclear speckle-like signals in PIK3R1-SH3. PIK3R2 has been shown to localise to the golgi apparatus in U-2OS cells (https://www.proteinatlas.org/cgi-bin/carddisp.pl?gene=PIK3R2/cell).In addition to golgi apparatus, nuclear and cytoplasmic staining of PIK3R2 have also been reported (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PIK3R2&keywords=PIK3R2#localization). In MDA-MB-231, PIK3R2

Previous studies showed that KIT protein localised to the cell membrane in HEK293 and HEL cells (https://www.proteinatlas.org/ENSG00000157404-KIT/cell). Staining of KIT protein have also been observed in the nucleus (https://www.genecards.org/cgi-bin/carddisp.pl?gene=KIT&keywords=KIT#localization). The staining presented in this work suggests that KIT-Y703 and KIT-pY721 expression were localised mainly to the cytoplasm in MDA-MB-231. Expression of KIT-cytoplasmic was mainly localised to the cytoplasm and strong speckle-like signals were observed in the nucleus in some MDA-MB-231 cells. However, the KIT staining observed in MDA-MB-231 in this work is, in general, very low. A few studies have demonstrated that KIT is lowly expressed in MDA-MB-231 [471, 472]. The intensity of IF staining is only a rough estimation of the protein expression level, and to quantify the expression level for KIT in MDA-MB-231, western blotting experiments need to be performed (discussed in Section 6.4).



Figure 6.4: Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in MDA-MB-231. Cells were fixed with iced cold methanol and stained with KIT/pKIT (green) and PIK3R1/PIK3R2 (red) antibodies. Nucleus was stained blue using DAPI. Merge panels show the co-distribution of KIT/pKIT, PIK3R1/2 and nucleus. Aurora/TPX2 was used as a positive control. Scale bar =  $10 \ \mu$ m.

Protein expression patterns were then examined in two Basal-like cell lines, MDA-MB-468 and LGI1T. Similar to Claudin-low cell line, Basal-like cell lines also represent TNBC characteristics (ER-, PR- and Her2-). However, Basal-like cell lines exhibit high expression of Ki-67 prognostic marker and down-regulation of cytokeratin 5 and cytokeratin 6. Results from previous studies showed that both KIT and PIK3R1 are expressed in MDA-MB-468 cell line [471–473]. As LGI1T is not a commercial cell line (isolated by Prof Valerie Speirs' group), the expression of KIT, PIK3R1 and PIK3R2 in this cell line is unknown. Staining for KIT/pKIT and PIK3R1/2 expression in the Basal-like phenotype cell lines MDA-MB-468 and LGI1T are shown in Figure 6.5 and Figure 6.6.

In MDA-MB-468, the expression of PIK3R1-SH2 was localised to cytoplasm with some strong dot-like signals. While PIK3R1-SH3 antibody is expected to bind to PIK3R1, and therefore expected to be localised to the same sub-cellular as the PIK3R1-SH2 staining , expression of PIK3R1-SH3 was predominantly localised to the nucleus. The nucleus staining of PIK3R1-SH3 could be associated with PI3K monomer dimers, or a result of experimental artefacts (discussed in Section 6.4.1). In MDA-MB-468, the PIK3R2-RhoGAP staining overlaps with the nuclear staining. Similar to the staining observed in MDA-MB-231, some more intense staining was also observed near the edge of nucleus. Again, this more intense staining can be associated with golgi apparatus (where PIK3R2 was observed in another cell line), however this assumption will need further validation.

Expression of KIT-Y703, KIT-pY721 and KIT-cytoplasmic in MDA-MB-468 were localised to cytoplasm, with some strong dot-like signals at the outer edge of cytoplasm. These strong dot-like signals could be associated with cell membrane (where the KIT expected cellular localisation is, information obtained from: https://www.genecards.org/cgi-bin/carddisp.pl? gene=KIT&keywords=KIT#localization). To examine whether these signals are membrane associated, a membrane marker or high-resolution light microscopy images would be required.

In the other Basal-like cell line, LGI1T, expression of PIK3R1-SH2 and PIK3R1-SH3 were localised to nucleus, with some strong cytoplasmic dot-like signals observed for PIK3R1-SH2. The strong dot-like signals in cytoplasm could be a result of unspecific binding, or detecting protein aggregation complex. PIK3R2-RhoGAb staining was seen in both nucleus and cytoplasm. Expression of KIT-Y703 and KIT-pY721 were localised to the cytoplasm. While the expression of KIT-cytoplasm is also localised to the cytoplasm, the expression intensity around the nucleus was slightly stronger than other regions of cytoplasm (signified by yellow arrows). This uneven expression intensity was seen in most LGI1T cells and suggests that KIT-cytoplasmic could be associated with specific subcellular compartments that are close to the nucleus, e.g. rough endoplasmic reticulum. However, this needs to be confirmed by future experiments.



Figure 6.5: Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in MDA-MB-468. Cells were fixed with paraformaldehyde and stained for immunofluorescence microscopy with KIT/pKIT (green) and PIK3R1/PIK3R2 (red) antibodies. Nucleus was stained blue using DAPI. Merge panels show the co-distribution of KIT/pKIT, PIK3R1/2 and nucleus in MDA-MB-468. Aurora/TPX2 was used as a positive control. Scale bar =  $10 \ \mu$ m.



Figure 6.6: Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in LGI1T. Cells were fixed with iced cold methanol and stained for immunofluorescence microscopy with KIT/pKIT (green) and PIK3R1/PIK3R2 (red) antibodies. Nucleus was stained blue using DAPI. Merge panels show the co-distribution of KIT/pKIT, PIK3R1/2 and nucleus. Merge panels show the co-distribution of KIT/pKIT, PIK3R1/2 and nucleus in LGI1T. Aurora/TPX2 was used as a positive control. Scale bar = 10  $\mu$ m.

Finally, to examine protein expression patterns in healthy breast cells, KIT/pKIT and PIK3R1/2 were also stained in a healthy breast cell line, HBL100 (Figure 6.7). In HBL100 cells, PIK3R1-SH2 and PIK3R1-SH3 were both localised to the nucleus and cytoplasm. Both nuclear and cytoplasmic staining of PIK3R1 have been reported by previous studies (https://www.genecards. org/cgi-bin/carddisp.pl?gene=PIK3R1#localization). Expression of PIK3R2-RhoGAP was localised predominantly to the nucleus with strong speckle signals observed. Similarly, nuclear staining of PIK3R2 have also been shown by previous studies (https://www.genecards. org/cgi-bin/carddisp.pl?gene=PIK3R2&keywords=PIK3R2#localization). Finally, expression of KIT-Y703, KIT-pY721 and KIT-cytoplasmic were all localised to the cytoplasm. The localisation of the protein epitopes in the four cell lines are summarised in Table 6.3.



Figure 6.7: Immunofluorescence images of KIT, pKIT, PIK3R1 and PIK3R2 in HBL100. Cells were fixed with paraformaldehyde and stained for immunofluorescence microscopy with KIT/pKIT (green) and PIK3R1/PIK3R2 (red) antibodies. Nucleus was stained blue using DAPI. Merge panels show the co-distribution of KIT/pKIT, PIK3R1/2 and nucleus in HBL100. Aurora/TPX2 was used as a positive control. Scale bar = 10  $\mu$ m.

tterns	
The pa	
lines.	
nt cell	
liffere	
is in e	
analys	
cence	
fuores	
munol	
the im	
from 1	
realed	
on rev	
alisati	ckets.
in loc	he brac
prote	ed in t
of the	describ
ımary	on are
3: Sun	calisati
able 6.5	the lo
Ĥ	of

	PIK3R1-SH2 (bs-PIK3R1)	PIK3R1-SH3 (nbs-PIK3R1)	PIK3R2- RhoGAP (nbs-PIK3R2)
MDA-MB-231 (claudin-low)	Cytoplasmic	Cytoplasmic, Nuclear (sparkle)	Cytoplasmic
MDA-MB-468 (Basal-like)	Cytoplasmic	Nuclear	Nuclear
LGI1T (Basal-like)	Nuclear	Nuclear	Cytoplasmic, Nuclear (sparkle)
HBL100 (healthy)	Cytoplasmic, Nuclear (diffuse)	Cytoplasmic, Nuclear (diffuse)	Nuclear
	KIT-Y703 (bs-KIT)	KIT-pY721 (bs-pKIT)	KIT-cytoplasmic (nbs-KIT)
MDA-MB-231 (claudin-low)	Cytoplasmic	Cytoplasmic	Cytoplasmic, Nuclear (sparkle, some cells)
MDA-MB-468 (Basal-like) LGI1T (Basal-like)	Cytoplasmic Cytoplasmic	Cytoplasmic Cytoplasmic	Cytoplasmic Cytoplasmic
$(\mathrm{HBL100(healthy)})$	Cytoplasmic	Cytoplasmic	Cytoplasmic

#### 6.3.2 In Situ Proximity Ligation Assay results

The next aim was to explore whether any colocalisation of PIK3R1/2 and KIT/pKIT can be detected by Proximity Ligation Assay (PLA). PLA is an assay that allows the identification of protein colocalisation. In particular, it generates a fluorescent signal when two proteins are within 40nm [474]. Unlike the dual immunofluorescence assay where colocalisation involved spatial overlapping of two different colours of signals, PLA produces one colour spot-like signal, which makes defining and quantifying colocalisation easier.

In the bs-PIK3R1+KIT experiment, two binding-site recognising antibodies (PIK3R1-SH2 and KIT-Y703) were considered. The experiment showed an significant (p<0.0001) amount of clear PLA signals in the three breast cancer cell lines (MDA-MB-231, MDA-MB-468 and LGI1T). In most cells, the bs-PIK3R1+KIT PLA signals are in a linear arrangement, suggesting they may be associated with cell membrane. The amount of bs-PIK3R1+KIT PLA dots observed from HBL100 was not significant compared to the negative control, suggesting that these signals are likely to be noise and do not represent protein colocalisation (Figure 6.8D).

While bs-PIK3R1-pKIT showed significant (p<0.0001 or p<0.05) amount of PLA signals in the three triple-negative breast cancer cell lines, the PLA signals observed for this antibody combination were fewer and weaker in comparisons to other antibody combinations (Figure 6.8A, B, C). Moreover, similar to the bs-PIK3R1+KIT experiment, bs-PIK3R1+pKIT did not show a significant amount of PLA signals in HBL100 (Figure 6.8D). This observation suggests that the PIK3R1-KIT/pKIT associations in triple-negative breast cancer cells (both claudin-low and Basal-like) might have different characteristics than those in healthy breast cells.

Finally, while *nbs-PIK3R1+KIT* and *nbs-PIK3R2+KIT* both contained antibody pairs that do not recognise the putative binding sites, they both showed a significant amount of PLA dots in all four cell lines. This suggests that while these regions are not putative binding sites, they are likely to be close (within 40nm) to each other, indicating either direct interactions or some level of physical associations (e.g. involve in a large protein complex). It is also worth noting that, different from PIK3R1-KIT/pKIT experiments where PLA signals are (likely to be) membraneassociated, the PLA signals from PIK3R2-KIT were observed in nucleus in MDA-MB-231 and HBL100 (Figure 6.8A, D). This suggests that PIK3R1 and PIK3R2 interact with KIT/pKIT in different cellular compartments and may have distinct roles in signal transduction. For example, PIK3R1 might be involved in the activation or a more upstream event of the signal transduction



pathway, whereas PIK3R2 might be involved in a more downstream signalling event.

Figure 6.8: **PLA analysis for the four antibody pairs in four cell lines.** (A) MDA-MB-231, (B) MDA-MB-468, (C) LGI1T and (D) HBL100 cells were fixed and stained with KIT/pKIT and PIK3R1/2 antibodies for PLA. Nucleus was stained blue using DAPI. Red signals indicate KIT/pKIT and PIK3R1/2 colocalisation. The left panel shows a typical negative control, where only single KIT antibody staining was used.



Figure 6.9: **PLA signal scatter plots.** The plots represent the distribution of the number of PLA signal/cell from (A) MDA-MB-231, (B) MDA-MB-468, (C) LGI1T and (D) HBL100 PLA analysis shown in Figure 6.8. Numbers above the negative controls indicate the average number of dots observed per cell. Data represent counts from three independent experiments \*\*\*\*p<0.0001; \*\*\*p<0.001; \*\*p<0.01; \*p<0.05.

### 6.4 Discussion and speculative research

In this chapter, four different pairs of antibodies that recognise different regions or forms of KIT, PIK3R1 or PIK3R2 were examined in four cell lines. Using the indirect dual immunofluorescence assay, the protein localisation in different cell lines were revealed. The immunofluorescence results suggested that some proteins, for example PIK3R1, localised to different sub-cellular compartments in different cell lines. Different protein localisations can be associated with different biological functions, and pathways involved. The protein localisation pattern observed in this work will be discussed in Section 6.4.1. While dual immunofluorescence assay was useful for identifying protein localisation, it was difficult to perform meaningful colocalisation analysis from the immunofluorescence images as some protein staining were weak or noisy. Therefore, PLA was used further to examine the colocalisation of the proteins of interest. While the PLA results demonstrated in this chapter generally supports the present of KIT-PIK3R1 and KIT-PIK3R2 PPIs, the results vary between cell lines and these differences in colocalisation characteristics will be discussed in Section 6.4.2. Finally, an additional experiment was planned to investigate the effect of KIT inhibitors on the protein colocalisation, however, due to the disruption caused by Covid-19 pandemic and laboratory closures, this experiment could not be carried furthure and therefore is considered speculatively. The original experimental design and expected outcome are discussed in Section 6.4.3.

#### 6.4.1 Cellular localisation of KIT, PIK3R1 and PIK3R2

In this work, the cellular localisation of KIT, PIK3R1 and PIK3R2 in breast cancer cell lines were examined using immunofluorescence assay. Three KIT antibodies recognising different epitopes of KIT were considered in this work, with one recognising the tyrosine 703 (KIY-Y703), one recognising the phosphorylated tyrosine 721 (KIT-pY721) and one recognising the cytoplasmic region of KIT (KIT-cytoplasmic). As a transmembrane receptor kinases protein, the expectation is that KIT would localised to the cell membrane. The membrane localisation of KIT has been previously seen in the adult mouse nervous system [475], spermatogonia cells of the adult testis [476], human ovarian follicles [477] and in breast epithelial cells [478]. In this work, the staining of KIT was predominantly localised to the cytoplasm. The cytoplasmic staining of KIT was also observed in spermatogonia cells of the adult testis [476], in addition to the membrane staining. It is hypothesised that this cytoplasmic staining of KIT represents a truncated form of KIT, which only contain the cytoplasmic domain and not the extracellular or transmembrane domains. In addition to adult testis, the truncated KIT, or known as soluble KIT, has also been observed in prostatic cancer [479], human umbilical vein endothelial cells [480] and hematopoietic cells [481, 482]. Different from full-length KIT, intracellular truncated KIT can regulate downstream signalling pathway without ligand-binding activation [483, 484]. Elevated expression of truncated KIT has been linked to increased activation in the Src pathway, and seen in prostate cancer [479]. While it is possible that the cytoplasmic staining of KIT observed in this work represents the existence of truncated KIT in breast cancer cell lines, most studies observed co-existence of both the full-length KIT (localised to the membrane) and truncated KIT (localised to the cytoplasm). Without observing the full-length KIT and the weak and noisy KIT staining presented in this work, it is more likely that the cytoplasmic staining observed is caused by unspecific binding of the antibodies.

In this work, expression of PI3K regulatory subunits were examined using two PIK3R1 antibodies and a PIK3R2 antibody. The two PIK3R1 antibodies recognise two different epitopes in PIK3R1, with one recognising the N-SH2 domain and one recognising the SH3 domain. The results presented in this thesis showed PIK3R1 can localised to both cytoplasm and nucleus. The results showed that PIK3R1 was predominantly localised to the cytoplasm in MDA-MB-231 and MDA-MB-468, and localised to the nucleus in LGI1T and HBL100 cells. This suggests that PIK3R1 is likely to be associated or is involved in nuclear processes such as transcriptional activities in LGI1T and HBL100. Different sub-cellular distribution was also observed for PIK3R2 staining, with cytoplasmic staining observed in MDA-MB-231 and LGI1T; and nuclear staining observed in MDA-MB-468, LGI1T and HBL100. As shown in the results section, some of the PIK3R2 staining had uneven intensity near the edge of DAPI nuclear staining. This could be associated with cellular compartments that are close to the nucleus such as golgi apparatus, where PIK3R2 was previously observed by other studies (https://www.proteinatlas.org/ENSG00000105647-PIK3R2/cell). However, this needs to be evaluated with golgi apparatus marker.

The nucleus localisation of PIK3R1 and PIK3R2 observed in this research could indicate the presence of "p110-independent PI3K regulatory subunits", a hypothesis proposed recently. p110 is the catalytic subunits of PI3K and it is hypothesised that some PI3K regulatory subunits can function without binding to the catalytic subunits (p110), independent of PI3K activity.

The p110-independent PI3K regulatory subunits are unstable and exist in a monomer-dimer equilibrium. This hypothesis was motivated by the finding from the previous biochemical study, where they suggested that PI3K regulatory subunits are likely to exist in two dimer forms, either complex with a PI3K catalytic subunits and exist as a heterodimer, or complex with another PI3K regulatory subunit and exit as a homodimer [485, 486]. While the concept of the PI3K regulatory homodimers is not fully understood [486], research has now shown that these dimers are independent of PI3K catalytic activity, and they can self-translocated to the nucleus or mediate nuclear translocation of other proteins such as XBP-1 in response to cell stress and other various growth factor signals [487–489]. Further research should investigate into whether PI3K regulatory subunits exits as homodimers in breast cancers, and, if they exist, what their involvements are in signalling transduction and how, when deregulated, this might lead be oncogenic.

In MDA-MD-468 and MDA-MB-231 cells, staining of PIK3R1-SH2 and PIK3R1-SH3 were observed in different cellular compartments. The inconsistent cellular localisation of the same protein indicates the possibility of antibody unspecific binding or experimental artefacts. While positive controls were performed during the immunofluorescence assay, the positive controls included in this work were only sufficient to suggest (a) there were no "human error" during the experiments and (b) the secondary antibodies and other regents were working. It should be noted that, the positive controls included in this work was not enough to demonstrate the legitimacy and the quality of the results. To demonstrate the staining results are reproducible and high-quality, positive controls need to show that the protein of interest are localised to the expected cellular compartment in a previously tested cell line. The expected cellular localisation of proteins can be accessed from previous literatures or online knowledge-based databank, such as The Human Protein Atlas (https://www.proteinatlas.org) or GeneCards (https://www.genecards.org/). In the case of this work, the most common cellular localisation for KIT is cell membrane, and without seeing KIT membrane-associated staining in a control cell line, it is difficult to examine whether the primary antibodies were working. While protocol optimisation took place prior the experiments, it was performed using MDA-MB-231, which is a breast cancer cell line that has been previously shown to have KIT expression alteration (low expression) by some studies [471, 472]. To establish the protocol, the healthy breast cell line HBL100 would be a better choice than MDA-MB-231, as HBL100 cells are more likely to represent the normal physiological cellular environment. If available, cell line HEK293 (a

commonly used mammalian cell line) would also be good cell line to use for establishing the protocol, as it is a popular cell line and much information, including expression of proteins, is shown and quantified by previous research. Protocol optimisations should carefully consider and identify the optimal condition of the following: fixation method, primary and secondary antibody incubation times and dilutions, and incubation temperature. To validate the IF results presented in this work, future work should optimise the protocol and include informative positive controls.

#### 6.4.2 KIT and PIK3R1/2 protein interactions and binding sites

While the results from immunofluorescence assay provide a useful indication of the protein localisation, some of the protein stainings were too weak or too noisy to perform meaningful colocalisation analysis. Therefore, to further investigate whether KIT colocalises with PIK3R1 and PIK3R2, proximity ligation assay (PLA) was carried out.

Whilst the KIT-PIK3R1/PIK3R2 protein-protein interactions have been previously demonstrated by several biochemical and biophysical approaches such as fluorescence polarisation spectroscopy [490] and communoprecipitation [330, 331], the PLA analysis presented in this chapter suggests that endogenous KIT colocalised with PIK3R1 and PIK3R2 in breast cancer cell lines. Using PLA, Liu et al. demonstrated that KIT-PIK3R1 interactions mostly localised to the cell membrane in HeLa cells [491]. The PLA results presented in this chapter are in agreement with Liu et al.'s observation, where we showed that most KIT-PIK3R1 PLA signals were in a linear arrangement at approximately where the cell membrane is. This is an expected outcome, as previous studies showed that KIT undergoes auto-phosphorylation on multiple tyrosine sites upon ligand binding and these phosphorylated tyrosine sites can recruit several downstream proteins to the cell membrane, including PIK3R1 [492–494]. The observed PLA signals are believed to represent protein signalling complexes. No significant difference in terms of PPI localisation was observed between breast cancer and healthy cell lines. However, this was only a first estimation and future work would need to stain the membrane with a fluorescent marker or include light microscopy images to confirm the localisation of endogenous KIT-PIK3R1 PPI.

In contract, KIT-PIK3R2 PPIs were localised to different cellular compartments in different cell lines. In particular, KIT-PIK3R2 were localised to the nucleus of the two non-Basal-like

cell lines (MDA-MB-231 and HBL100), and to the cytoplasm in the two Basal-like cell lines (MDA-MB-468 and LGI1T). PIK3R2 is often considered to be a similar protein to PIK3R1 and due to this reason, its specific biological functions and roles have been over-looked. The distinct PPI localisation observation here suggest that PIK3R1 and PIK3R2 may have different functions. In cancers, a growing number of studies suggest that PIK3R1 and PIK3R2 have different, and possibility, opposite functions, in cancer progression [463, 495, 496]. While the KIT-PIK3R2 interaction has been previously demonstrated in vitro [331], its cellular localisation remains unclear. Due to the high sequence similarity to PIK3R1, KIT-PIK3R2 interactions are assumed to occur via a similar mechanism to PIK3R1 and expected to localise to cell membrane. However, the nuclear localisation of KIT-PIK3R2 observed in this thesis disagrees with this assumption. Kumar et al. showed that PIK3R2, and not PIK3R1, is crucial for mediating the PI3K beta (PIK3R2/PIK3CB heterodimer) transport across the nuclear envelope and enter/exit the nuclear [497]. Importantly, this nuclear translocation of PIK3R2/PIK3CB heterodimer has been shown to regulate cell viability in multiple cell lines. A Recent study showed that PIK3R2 can bind to mutated PIK3CA and assist nuclear translocation of the complex in tumours [498]. While no evidence suggest that the PIK3R2-PIK3CB/A heterodimer interacts with KIT in the nucleus, it is not entirely impossible. Especially, given the high sequence similarity of the SH2 domains in PIK3R1 and PIK3R2, if KIT (or truncated KIT) can also translocate to the nucleus, it is likely that KIT and PIK3R2 can co-localised in the nucleus. However, this is a very preliminary assumption and further work need to take place to examine (a) whether KIT translocates to cell nucleus, (b) if they do, whether KIT interact with PIK3R2 in the cell nucleus, (c) how the nuclear PIK3R2-KIT complexation affect cell survival and tumorigenesis, the two processes that have been previously linked to PIK3R2-PIK3CB/A nuclear translocation [497, 498], and finally (d) whether PIK3R2-KIT PPIs are disturbed in Basal-like cell lines.

Differences were also observed in the number of PLA signals between different antibody pairs and cell lines. The four pairs of antibodies considered in this work were either raised to recognise putative binding sites of our PPIs of interest or were raised to recognise regions away from the binding sites. The PLA results showed that, on average, non-binding site (nbs) antibodies gave more PLA signals per cell compared to the binding site (bs) antibodies. This could be due to the availability of the exposed epitopes in the proteins (Figure 6.10). In a binary complexation scenario, two proteins can be in three forms: un-complexed, transient or complexed. In an uncomplex form, two proteins are distant from each other, and although in this stage antibodies, both binding site antibody (bs-ab) and non-binding site antibody (nbs-ab), are able to bind to the epitopes, they could be too far away (> 40nm) from each other to produce PLA signals. Most of the protein-protein interactions are thought to be transient in nature, meaning that their association and disassociation are continuous processes. This intermediate state between complete un-association and complete complexation is referred to as transient form. In this form, the interactions (attractions and repulsions) between the two proteins are weak, and proteins will start to disassociate from short-lived complexes. In the transient form, the epitopes in the proteins are still exposed, therefore, antibodies, both bs-antibodies and nbs-antibodies, would be able to bind to the corresponding protein targets. Depending on the distance between the epitopes and the attached antibodies, PLA signals might be produced if the distance is smaller than 40nm. In the third form, the two proteins are complexed, meaning the binding sites from both proteins are occupied and not exposed at the protein surface. In this form, bs-antibodies would not be able to bind to corresponding epitopes on their protein targets, therefore, although the proteins are within 40nm, no PLA signals will be observed. However, since non binding site regions are not involved in the complexation, they are likely to be exposed at the protein surface, allowing some antibody binding. Due to this reason, nbs-antibodies are likely to give positive PLA signals. This could explain why bs-antibodies, especially KITpY721 to PIK3R1-SH2 experiments, gave the least amount of PLA signals, whereas the two non binding site antibody combinations showed more PLA signals. The PLA results also showed that bs-PIK3R1+KIT and bs-PIK3R1+pKIT gave a significant amount of PLA signals in TNBC cells but not in healthy breast cells. This could suggest that the PIK3R1-KIT PPIs in triple negative breast cancer are less stable (high  $K_{off}$ ), and there are more PIK3R1 and KIT proteins in the transient form in triple negative breast cancer than there are in normal breast cells. This could be a disruption caused by abnormal aggregation of KIT in triple negative breast cancer cells. The immunofluorescence staining presented in this work shows that the expression pattern of the two binding site epitopes in KIT (KIT-Y703 and KIT-pY721) tend be sparkled in the three triple negative breast cancer cell lines and diffused in the healthy HBL100 cells, indicating that KIT may form some levels of abnormal aggregations in triple negative breast cancer cells. It has been demonstrated that the formation of abnormal protein aggregation can disrupt functional protein complexes and lead to toxicity [499]. Moreover, studies have suggested that proteins with intrinsically disordered regions are prone to aggregate formation [500, 501]. With both KIT-Y703 and KIT-pY721 located at the intrinsically disordered regions of KIT, the results presented in this chapter highlight an interesting observation that has not been previously reported, adding a valuable avenue for further research. Future work can use microscopic techniques to determine whether KIT form abnormal aggregation in TNBC, and if the aggregation is present, kinetic methods can be used to monitor the aggregation process in real time [502]. In additional to experimental methods, computational methods (e.g. molecular dynamics simulations to estimate a protein's amyloidogenicity) can provide additional structural, dynamic, and mechanistic insights for investigating protein aggregation [503].

It should be pointed out that the scenarios shown in Figure 6.10 was based on the assumption that the putative binding sites described in the previous chapters were genuine, and the interaction between KIT and PI3K regulatory subunit was binary. While the PLA results shown in this chapter confirmed the KIT-PIK3R1 and KIT-PIK3R2 colocalisation and suggested PPIs between KIT and PI3K regulatory subunits in breast cancerous cells, further biochemical analyses are required to explore their molecular mechanisms of binding.


Figure 6.10: Schematic illustration the concept of antibody binding and epitopes exposure in complexed, transient and complexed protein-protein interaction (PPI) form. The three forms are reversible. (A) and (B) show the binding site interacting antibodies (bs-ab) and non-binding site interacting antibodies (nbs-ab) in the three PPI forms, respectively. KIT proteins are indicated with wavy lines. PIK3R1/PIK3R2 proteins are indicated with a blue oval. Phosphate groups are indicated with P circles.

#### 6.4.3 Speculative research and future work

Once PLA has confirmed the colocalisation of KIT-PIK3R1 and KIT-PIK3R2, the next aim was to examine the effect of KIT inhibitors on the protein expression level and protein colocalisation in the four cell lines and predict whether Basal-like patients are likely to benefit from KIT inhibitors. However, as a result of Covid-19 and laboratory closure, the experiments could not be carried out and the following had to be considered speculatively.

The first step of this experiment would involve identifying commercially available KIT inhibitors and to understand their characteristics in order to select suitable inhibitors for the experiments. A total number of 38 commercially available KIT inhibitors can be identified from Selleckchem (https://www.selleckchem.com/c-Kit.html). In which 17 of them have half-maximal inhibitory concentration  $(IC_{50})$  value for KIT below 10 nM (Figure 6.11).  $IC_{50}$  is the concentration of an inhibitor at which the target receptor or enzyme activity is 50% inhibited. Therefore, the lower the  $IC_{50}$ , the more potent the inhibitor is. Low  $IC_{50}$  also reduces the chance of undesired off-target binding. The sensitivities of the 17 inhibitors with KIT  $IC_{50}$  lower than 10 nM in cancer cell lines were further evaluated using Genomics of Drug Sensitivity in Cancer (Release 8.3, https://www.cancerrxgene.org), which screened more than 1000 cancer cell lines with a wide range of anti-cancer therapeutics [504]. This database showed that out of the 17 low  $IC_{50}$  KIT inhibitors, Axitinib and Amuvatinib were significantly ( $IC_{50}$  z-score > 2.0) sensitive to MDA-MB-468, which is a Basal-like cell line used in our experiments. These two inhibitors would be a good starting point for exploring the effect of KIT inhibitors in Basal-like breast cancer. In addition to the two inhibitors, other research has also highlighted the pharmaceutical potential of other KIT inhibitors such as Imatinib and Dasatinib in Basal-like breast cancer [505 - 507].



Figure 6.11: Schematic diagram of proposed work. The process of identifying KIT inhibitors are shown in grey boxes. The proposed experiments will explore the effect of appropriate KIT inhibitors on cell behaviours (green) and endogenous KIT/pKIT and PIK3R1/2 (blue). WB: western blotting; IF: immunofluorescence; PLA: proximity ligation assay.

The next aim of the experiment would be to explore the importance of KIT in Basal-like breast cancer by examining the effect of the KIT inhibitors (e.g. Axitinib) in the four breast cell lines used in the previous experiments. To achieve this, the experiments would involve treating the cells with different doses of inhibitors for different time course and monitor the cell behaviours (e.g. proliferation, apoptosis or cell viability) accordingly. To examine whether KIT inhibitor has an effect on the expression level of KIT, PIK3R1, PIK3R2 and other downstream proteins (e.g. AKT, extracellular signal-regulated kinases (ERK), STAT proteins), western blotting experiments should be carried out to quantify the expression level of the proteins with and without treatment. In addition, to examine whether the KIT inhibitor has an effect on the KIT-PIK3R1 and KIT-PIK3R2 PPIs, their protein localisation and colocalisation should also be examined in the presence of the inhibitor using dual immunofluorescence and PLA, respectively. The IF and PLA results should be compared to the results presented in this chapter to explore the potential mechanism of the KIT inhibitors in different breast cell lines.

While the experiments could not be carried out, the expected outcome is that KIT inhibitor is

likely to change (most likely reduce) the KIT expression level at a relatively low concentration. The inhibition of KIT is likely to interrupt KIT-mediating PPIs, including KIT-PIK3R1 and KIT-PIK3R2. Therefore, it is expected that, with treatment, there will be a reduction of PLA signals in cells. The PLA analysis would help clarity whether this reduction is significant and whether this reduction differs in different cell lines. To further examine whether the inhibitor disrupts the PPIs of interest, although not part of the proposed plan, more sophisticated assays such as immunoprecipitation of endogenous protein could be carried out too, to confirm the fidelity of the antibody binidng.

The disruption of KIT-mediating PPIs is likely to affect the proteins involved in the downstream signalling pathways, therefore the protein expression of PIK3R1, PIK3R2 and other downstream proteins (e.g. AKT, mTOR) are expected to change (either upregulated or downregulated). KIT receptor downstream signal transduction pathways includes Ras/Erk signal transduction pathway (regulates cell differentiation proliferation and migration), PI3K/AKT pathway (associated with anglogenesis), PLC- $\gamma$  pathway (regulates proliferation and survival), Src pathways (regulates cell survival, proliferation, motility, migration, invasion and anglogenesis). Interfering KIT and its activation will likely to affect the the biological processes controlled by its downstream signaling pathway. Several studies have demonstrated that KIT is overexpressed in Basal-like and might serve as a potential pharmaceutical target [314, 315, 508]. While it is still unclear how inhibiting KIT would affect Basal-like cells and downstream proteins, previous studies have demonstrated promising results in other cancers including gastrointestinal stromal tumor, acute myeloid leukemia, melanoma [509]. The experiments described in this chapter may provide an early insight into understanding the roles of KIT in Basal-like and uncover the dynamic between KIT and proteins involved in signal transduction. Previous study has demostarting that inhibiting KIT using STI571 in metastatic gastrointestinal Stromal Tumor patients has been shown to be effective. Ikeda et al. showed that ABT-869 can inhibit proliferation in Ewing Sarcoma cells by blocking c-kit receptor signaling pathway [510]. There are also a number of studies demonstrating that blocking stem cell factor (SCF)/c-Kit pathway by imatinib lead to repressed solid tumors growth and metastasis [511–514]. However, a more recent study showed that targeting KIt have no or very limited affect on cell proliferation and radioresponse in several non-small-cell lung carcinoma or prostate cancer cells [515]. The controversial findings suggest that the effect of interfering KIT signaling pathway can be context and cell line depending. It is also important to recognise that, as with many kinase inhibitors,

KIT inhibitors as a single target therapy might not give rise to significant activities in Basal-like. Further strategies should explore other RTKs and their relationship with KIT in Basal-like for developing rational combination therapies.

### 6.5 Key findings

- PIK3R1 and PIK3R2 have distinct cellular expression pattern (localised to different cellular compartments), suggesting distinct biological roles.
- KIT colocalised with PIK3R1 and PIK3R2 in breast cancer cells.
- KIT inhibitors such as Axitinib and Amuvatinib or KIT ligand stem cell factor would be suitable candidates for manipulating KIT-PIK3R1 and KIT-PIK3R2 PPIs and exploring the consequent effects in cell lines.

## Chapter 7

## **Discussion and Future work**

The methods developed in this thesis demonstrates a novel computational drug target identification pipeline. The approach integrates multidisciplinary techniques, including unsupervised clustering, network modelling, systems biology approaches, molecular modelling and structural biochemistry. Integrating transcriptional data with proteomic data, this strategy explored a number of critical SH2-mediating PPIs as potential therapeutic targets for Basal-like breast cancer, generating a focussed and testable hypothesis for further research.

As mentioned earlier, most Basal-like breast cancer lack expression of ER, PR and Her2, making them unsuitable and unresponsive to current clinical targeted therapies that target these proteins. Therefore, to address this clinical challenge, the primary aim of this thesis was to identify therapeutic targets to treat Basal-like breast cancer. The first analysis presented in this thesis uncovered the underlying transcriptional characteristics associated with Basal-like breast cancer. In this analysis, unsupervised Bayesian Hierarchical Clustering (BHC) was applied to reveal a natural stratification of (i) breast cancer patients according to their transcriptional expression profiles and; (ii) transcriptional variables according to the patients' organisation. The clustering results presented in this thesis demonstrated that Basal-like breast cancer has distinct transcriptional expression profile from Luminal-type breast cancer, suggesting they might arise from distinct cells within the tissue (cell-of-origin hypothesis). However, a fraction of Basal-like breast cancer has similar transcriptional expression pattern to Her2 patients, indicating that the differences between Basal-like and Her2 subtype are more subtle and might present more similar phenotypes. This might suggest that Basal-like and Her2 subtype have the same cell of origin and the two subtypes are primarily determined by a series of genetic and epigenetic events. The Her2-like BLBC tumour observed in this study might reflect a subset of earlystage tumour cells that have not evolved into a distinct subtype or a novel subset of tumour that display both Basal-like and Her2 characteristics. While the clustering results presented in this thesis only provide an early insight into the phenotypic heterogeneity of breast cancer, the ambiguity observed in this study raised the interesting possibility that the Basal-like and Her2 subtypes share some homology and may not represent distinct molecular subtypes.

The BHC results also revealed 22 Basal Candidate Genes (BCGs) that are highly expressed solely in Basal-like breast cancer. With the primary aim of this project being to identify therapeutic targets for Basal-like breast cancer, the natural progression of this study was to prioritise the 22 BCGs in order to perform a more focused and sophisticated analysis at the protein level. However, it is possible that these 22 BCGs can provide more insights especially in the area of pharmaceutical and clinical research. For example, they might be valuable to pinpoint candidate biomarkers or gene signatures for diagnostics, prognostics or predictive purposes. To assess whether any of the 22 BCGs could be a clinically valuable biomarker, future work should explore how the expression of these genes correlates with patients' characteristics (e.g. recurrence rate, survival rate, response to treatment).

The partial correlation networks developed in this work revealed topology characteristics of the 22 BCGs, allowing identification of critical positions and relationships within the network. Among the 22 BCGs, LCK and STAT1 were found to be the most connected, representing critical candidate genes. The importance of LCK and STAT1 in breast cancer has also been reported by other studies [253, 516–519]. As LCK and STAT1 are known to be involved in T-cell proliferation [520] and IFN $\gamma$ -related defence response [244] respectively, it is possible that the LCK and STAT1 expression observed in the study are linked to the presence of lymphocyte infiltration within the tumour, indicating that Basal-like tumours are heterogeneous. This hypothesis was in line with the results reported by a recent study done by Thompson *et al.*, where they showed high lymphocytic infiltrate was observed in triple negative breast cancer, a clinical breast cancer subtype that highly overlaps with Basal-like breast cancer [521]. While results presented in this thesis provide preliminary evidence of the existence of lymphocyte infiltrates in Basal-like breast cancer, future work should explore the roles and compositions of the immune environment, especially lymphocytic infiltrates, in Basal-like tumours.

In this study, the critical correlation between LCK and STAT1 mRNA expression was further

explored at the protein level. The protein-protein interactions (PPIs) that are associated with or mediate between LCK and STAT1 were showed in a PPI network. This PPI network was developed based on the reported data and the available literature at the time (June 2018). Such data/knowledge-based approach often reflects only a fraction of the "true system" and can be biased. For example, popular proteins (such as kinases or onco-proteins) will likely to receive more research attention than those poorly characterised proteins, resulting in a varying amount of data and literature. However, this does not devalue the 12 critical PPIs identified from this work, instead, it suggests that apart from the 12 PPIs identified from this work, there might be other critical PPIs that mediate between LCK and STAT1 that have not yet been reported or discovered. Repeating the strategy described in Chapter 4 in the future would reveal an updated vision of the LCK and STAT1 PPI network.

Another consideration when developing PPI networks from genomic data is the mRNA-protein translation rate. In this work, the mRNA expression of LCK and STAT1 were assumed to be the first proxy of protein expression and their strong association was assumed to be reflected at the protein level. While Wilhelm *et al.* demonstrated that mRNA expression typically correlates well with protein expression and can be a useful indication to predict the protein abundance level [522], Fortelny *et al.* showed that many mRNA-protein expression correlations have been overestimated and argues that mRNA expression is, in general, a poor predictor of protein abundance levels [344]. While whether mRNA expression is a good indicator for estimating protein expression remain controversial, both Wilhelm *et al.* and Fortelny *et al.* agreed that gene-protein translation is a highly complex process and is affected by multiple translational and post-translational factors. Nevertheless, before we are able to sample a proteome comprehensively (all proteins, all the time), or in this case, where no proteomic data were available, mRNA expression still provides a useful first proxy for estimating the abundance of proteins.

From the LCK and STAT1 PPI partial correlation network, 12 critical PPIs were revealed. Among these, the protein interactions between GAB1, KIT and PI3K regulatory subunits were explored further in this thesis. Molecular docking performed in Chapter 5 model the interactions between short SLiM-containing peptides from KIT/GAB1 and SH2 domains from PI3K regulatory subunits at the atomic level, revealing specific molecular recognition features of the peptide-SH2 complex. The peptide docking results presented in this thesis indicate that the four phosphorylated-tyrosine (pY) peptide extracted from KIT and GAB1 bind to most N-SH2 domain and C-SH2 domain in PI3K regulatory subunits, and no obvious preferences were observed toward a specific SH2 domain. Due to the high sequence similarity between N-SH2 and C-SH2, it is likely that apart from the geometric and chemical complementarity between the proteins of interest, the binding specificities of N-SH2 and C-SH2 are achieved by other factors, such as cellular localisation. It is also possible that these PPIs are transient by nature (regularly associate and disassociate rapidly) and need more sophisticated analyses to reveal their specificity. The specificity of transient PPIs, especially those are mediated by SLiM-containing peptides, could also be determined by residues adjacent to SLiMs [523, 524]. While proteome-wide peptide array experiments done by Tinti et al. suggested that SH2 domains, in general (not specific to PI3K regulatory subunits), discriminate their binding partners by making favourable contacts with residues from position +1 up to +4 in the peptide [525]. The result shown in Chapter 5 suggested that in PI3K SH2 domains, favourable contacts were made with both residues Nand C-terminal to the pY. This observation should be a motivation for future work to explore different length of peptides and how each residue at each position contribute (stabilise or destabilise) to the binding. Understanding the involvement of contextual sequences in the peptide would allow us to gain insight into the SH2 binding specificity and inform the development of isoform-selective inhibitors or therapeutic molecules. Furthermore, as the peptide docking carried out in this study was performed in vacuo, it would be interesting to explore the effect of water molecules in PI3K SH2 domain binding. This could be achieved by performing an equivalent set of docking simulations with the inclusion of surface-bound and conserved water molecules.

Using cellular assays, the expression pattern of endogenous KIT and PI3K regulatory subunits were explored in four different breast cell lines. The immunofluorescence results presented in this work show that KIT was predominantly localised to cytoplasm and PIK3R1 and PIK3R2 localised to different cellular compartment in different cell lines. For example, PIK3R1 N-SH2 domain expression was only cytoplasmic in MBA-MB-231 and MDA-MB-468 but both cytoplasmic and nuclear in LGI1T and HBL100. Nuclear staining was also observed in PIK3R2 in MDA-MB-468 and HBL100. The nuclear staining of PIK3R1/2 observed in this study suggested that PIK3R1/2 may be able to translocate into the cell nucleus and play a role in transcriptional regulation with other transcriptional factors or other nuclear-associated biological functions. It is important to note that, the nuclear staining of PIK3R1/2 were observed in both cancerous and healthy breast cell lines, and as first evidence, the results here are not suggesting there is an association between PIK3R1/2 nuclear localisation and Basal-like breast cancer. Nevertheless, it would still be interesting to explore whether PIK3R1/2 nuclear localisation is specific to breast cell lines, as well as why and how PIK3R1/2 translocate to the cell nucleus.

Apart from exploring the expression pattern, the experimental work carried out in this study also aimed to explore whether the critical PPIs identified from the PPI network present in Basal-like cell lines. Due to the availability of commercial antibodies, only two pairs of protein colocalisation, KIT-PIK3R1 and KIT-PIK3R2, were explored. However, if antibody of interest become available in the future, or one is able to synthesise them, the strategy described in Chapter 6 can be adopted to explore the existence of other critical PPIs in breast cancer cell lines. The PLA results shown in Chapter 6 suggested that endogenous KIT colocalised with PIK3R1 and PIK3R2 in the three breast cancer cell lines examined. Interestingly, it is observed that, on average, non-binding site (nbs) antibodies gave more PLA signals per cell compared to the binding site (bs) antibodies. This could be due to the availability of the exposed epitopes in the proteins. For example, in PLA experiments where bs-antibodies are considered, the putative binding sites might be involved in the corresponding PPI (occupied) and not available for the bs-antibodies to bind to, therefore, less PLA signals. On the other hand, epitopes that are not involved in the PPIs are more likely to be recognised by antibodies. Nevertheless, the PLA signals in both non-binding site and binding site experiments indicate that KIT and PIK3R1/2are in close proximity (< 40 nm) in the four cell lines examined in this study, suggesting that KIT does interact with PIK3R1/2 in Basal-like breast cancer. As a drug target identification project, the natural progression of this work would be to explore how interfering the KIT-PIK3R1, and KIT-PIK3R2 PPIs would affect cellular behaviour. However, as a result of Covid-19, the lab was forced to close, therefore, the experiments could not be carried out and had to be considered speculatively. The proposed work described in Chapter 6 Section 6.4.3 aims to explore the effect of KIT inhibitors; however, it would also be interesting to explore how the cells behave in response to KIT over-activation. This could be achieved by treating cells with stem cell factor, a cytokine that binds to and activates KIT receptor. An understanding of the effect of inhibitors or ligands on breast cancer cell behaviour will indicate whether these PPIs are likely to be a valuable drug target. If they are, the corresponding molecular recognition features described in Chapter 5 will provide an informative starting point for developing inhibitors that target these PPIs. On the other hand, if they are not valuable drug targets, the strategy described in Chapter 5 and Chapter 6 could be repeated to explore the rest of the critical PPIs (e.g.

KIT-RASA1) identified from the PPI network.

Integrating open access multi-omics data and system biology approaches, the strategy presented in this thesis provide a cost-effective approach for revealing critical PPIs that may serve as a therapeutic target in Basal-like breast cancer. While the analyses presented in this thesis are based on TCGA and METABRIC, the two most comprehensive breast cancer genomics data sets (at the time of analysis). It should be acknowledged that in the future, there is the likelihood that there will be additional genomic projects with a larger cohort or more comprehensive information. Therefore, future research may look to repeat the analysis with newer datasets. If further analyses revealed similar results to this work, it would confirm and strengthen the findings presented in this thesis. However, it is also likely that the new analyses would reveal novel PPI candidates for targeting. In this case, the new potential targets would need to be validated experimentally to examine whether they are directly involved in Basal-like, and whether modulating them is likely to have a therapeutic effect.

In conclusion, this study has integrated bioinformatics, system biology and structural biochemistry approaches to reveal critical PPIs in Basal-like breast cancer, providing a valuable stepping stone for future drug discovery research. In this study, 12 critical PPIs were identified from the genomic data, and among them, two were validated experimentally. The results showed that KIT-PIK3R1 and KIT-PIK3R2 PPIs are present in Basal-like cells, presenting potential PPI drug targets for Basal-like breast cancer. Furthermore, with most data and modelling tools used in this thesis being free to academics or open access, the approach developed in this work represents a modern and cost-effective approach for early drug target discovery research. Finally, it should also be pointed out that the method is not restricted to breast cancer and can be easily adapted to study other cancers or genetic diseases.

# References

- Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S. & Institute, U. N. H. G. R. A vision for the future of genomics research. *Nature*. 422, 835–847 (2003).
- Brittain, H. K., Scott, R. & Thomas, E. The rise of the genome and personalised medicine. Clin Med (Lond) 17, 545–551 (2017).
- Milioli, H. H., Tishchenko, I., Riveros, C., Berretta, R. & Moscato, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Medical Genomics* 10, 19 (2017).
- Roulot, A., Hequet, D., Guinebretiere, J. M., Vincent-Salomon, A., Lerebours, F., Dubot,
   C. & Rouzier, R. Tumoral heterogeneity of breast cancer. Annales de Biologie Clinique (Paris) 74, 653–660 (2016).
- 5. World Health Organization. WHO report on cancer: setting priorities, investing wisely and providing care for all 149 (World Health Organization, 2020).
- Cancer Research UK. Breast Cancer Statistics. https://www.cancerresearchuk.org/ health-professional/cancer-statistics-for-the-uk. Accessed 26 November 2020. 2018.
- 7. Office for National Statistics. Cancer registration statistics, England: 2017 https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2017. Accessed 26 November 2020. 2017.

- American Cancer Society. Breast Cancer Facts and Figures 2019-2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf. Accessed 26 November 2020. 2020.
- Mehrgou, A. & Akouchekian, M. The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. *The Medical Journal of the Islamic Republic of Iran.* 30, 369 (2016).
- Coles, C., Condie, A., Chetty, U., Steel, C. M., Evans, H. J. & Prosser, J. p53 Mutations in Breast Cancer. *Cancer Research.* 52 (1992).
- Lynch, E. D., Ostermeyer, E. A., Lee, M. K., Arena, J. F., Ji, H., Dann, J., Swisshelm, K., Suchard, D., MacLeod, P. M., Kvinnsland, S., Gjertsen, B. T., Heimdal, K., Lubs, H., Moller, P. & King, M. C. Inherited mutations in PTEN that are associated with breast cancer, cowden disease, and juvenile polyposis. *American Journal of Human Genetics*. 61, 1254–1260 (1997).
- Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkas, K., Roberts, J., Lee, A., Subramanian, D., De, L. K., Fostira, F., Tomiak, E., Neuhausen, S. L., Teo, Z. L., Khan, S., Aittomaki, K., Moilanen, J. S., Turnbull, C., Seal, S., Mannermaa, A., Kallioniemi, A., Lindeman, G. J., Buys, S. S., Andrulis, I. L., Radice, P., Tondini, C., Manoukian, S., Toland, A. E., Miron, P., Weitzel, J. N., Domchek, S. M., Poppe, B., Claes, K. B., Yannoukakos, D., Concannon, P., Bernstein, J. L., James, P. A., Easton, D. F., Goldgar, D. E., Hopper, J. L., Rahman, N., Peterlongo, P., Nevanlinna, H., King, M. C., Couch, F. J., Southey, M. C., Winqvist, R., Foulkes, W. D. & Tischkowitz, M. Breast-cancer risk in families with mutations in PALB2. *The New England Journal of Medicine.* 371, 497–506 (2014).
- Goldgar, D. E., Healey, S., Dowty, J. G., Da, S. L., Chen, X., Spurdle, A. B., Terry, M. B., Daly, M. J., Buys, S. M., Southey, M. C., Andrulis, I., John, E. M., Khanna, K. K., Hopper, J. L., Oefner, P. J., Lakhani, S. & Chenevix-Trench, G. Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Research.* 13, R73 (2011).

- Corso, G., Intra, M., Trentin, C., Veronesi, P. & Galimberti, V. CDH1 germline mutations and hereditary lobular breast cancer. *Familial Cancer.* 15, 215–219 (2016).
- Apostolou, P. & Papasotiriou, I. Current perspectives on CHEK2 mutations in breast cancer. Breast Cancer (Dove Med Press). 9, 331–335 (2017).
- Brewer, H. R., Jones, M. E., Schoemaker, M. J., Ashworth, A. & Swerdlow, A. J. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Research and Treatment.* 165, 193–200 (2017).
- Brown, K. F., Rumgay, H., Dunlop, C., Ryan, M., Quartly, F., Cox, A., Deas, A., Elliss-Brookes, L., Gavin, A., Hounsome, L., Huws, D., Ormiston-Smith, N., Shelton, J., White, C. & Parkin, D. M. The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *British Journal of Cancer.* 118, 1130–1141 (2018).
- Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao,
   P. P. & Zhu, H. P. Risk Factors and Preventions of Breast Cancer. *International Journal* of Biological Sciences. 13, 1387–1397 (2017).
- Soerjomataram, I., Louwman, M. W. J. & Ribot, J. G. An overview of prognostic factors for long-term survivors of breast cancer. *Breast Cancer Res Treat* 107, 309–330 (2008).
- NICE. National Institute for Health and Care Excellence (NICE) June 2018 https: //www.nice.org.uk/guidance/ng101. Accessed 28 March 2021. 2018.
- Patanaphan, V., Salazar, O. M. & R., R. Breast cancer: metastatic patterns and their prognosis. Southern Medical Journal 81, 109–12 (1988).
- TNM Classification of Malignant Tumours 8th edition (eds Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C.) (Wiley-Blackwell, 2016).
- Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Research and Treatment. 22, 207–219 (1992).

- 24. Fong, Y., Evans, J., Brook, D., Kenkre, J., Jarvis, P. & Gower-Thomas, K. The Nottingham Prognostic Index: five- and ten-year data for all-cause survival within a screened population. Annals of the Royal College of Surgeons of England. 97, 137–139 (2015).
- 25. Breast. In: AJCC Cancer Staging Manual. 8th ed. Springer (2017).
- Anderson, W. F., Chatterjee, N., Ershler, W. B. & Brawley, O. W. Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. Breast Cancer Research and Treatment 76, 27–36 (2002).
- Yu, K. D., Wu, J., Shen, Z. Z. & Shao, Z. M. Hazard of breast cancer-specific mortality among women with estrogen receptor-positive breast cancer after five years from diagnosis: implication for extended endocrine therapy. *The Journal of Clinical Endocrinology* and Metabolism. 97, E2201–2209 (2012).
- Truong, P. T., Bernstein, V., Wai, E., Chua, B., Speers, C. & Olivotto, I. A. Age-related variations in the use of axillary dissection: a survival analysis of 8038 women with T1-ST2 breast cancer. *International Journal of Radiation Oncology, Biology, Physics.* 54, 794–803 (2002).
- Bentzon, N., During, M., Rasmussen, B. B., Mouridsen, H. & Kroman, N. Prognostic effect of estrogen receptor status across age in primary breast cancer. *International Journal of Cancer.* 122, 1089–1094 (2008).
- Sopik, V., Sun, P. & Narod, S. A. The prognostic effect of estrogen receptor status differs for younger versus older breast cancer patients. *Breast Cancer Research and Treatment*. 165, 391–402 (2017).
- Jayasekara, H., MacInnis, R. J., Chamberlain, J. A., Dite, G. S., Leoce, N. M., Dowty, J. G., Bickerstaffe, A., Win, A. K., Milne, R. L., Giles, G. G., Terry, M. B., Eccles, D. M., Southey, M. C. & Hopper, J. L. Mortality after breast cancer as a function of time since diagnosis by estrogen receptor status and age at diagnosis. *International Journal of Cancer.* 145, 3207–3217 (2019).

- 32. Costa, R. L. B. & Czerniecki, B. J. Clinical development of immunotherapies for HER2+ breast cancer: a review of HER2-directed monoclonal antibodies and beyond. NPJ Breast Cancer. 6 (2020).
- 33. Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A., Lickley, L. A., Rawlinson, E., Sun, P. & Narod, S. A. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research.* 13, 4429–4434 (2007).
- Liedtke, C., Mazouni, C., Hess, K. R., Andre, F., Tordai, A., Mejia, J. A., Symmans, W. F., Gonzalez-Angulo, A. M., Hennessy, B., Green, M., Cristofanilli, M., Hortobagyi, G. N. & Pusztai, L. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *Journal of Clinical Oncology.* 26, 1275–1281 (2008).
- 35. Lin, N. U., Vanderplas, A., Hughes, M. E., Theriault, R. L., Edge, S. B., Wong, Y. N., Blayney, D. W., Niland, J. C., Winer, E. P. & Weeks, J. C. Clinicopathologic features, patterns of recurrence, and survival among women with triple-negative breast cancer in the National Comprehensive Cancer Network. *Cancer.* **118**, 5463–5472 (2012).
- Jr Goncalves, H., Guerra, M. R., Cintra, J. D., Fayer, V. A., Brum, I. V. & Bustamante Teixeira, M. T. Survival Study of Triple-Negative and Non-Triple-Negative Breast Cancer in a Brazilian Cohort. *Clinical Medicine Insights: Oncology.* 12, 1179554918790563 (2018).
- 37. Reddy, S. M., Barcenas, C. H., Sinha, A. K., Hsu, L., Moulder, S. L., Tripathy, D., Hortobagyi, G. N. & Valero, V. Long-term survival outcomes of triple-receptor negative breast cancer survivors who are disease free at 5 years and relationship with low hormone receptor positivity. *Breast Cancer Research.* **118**, 17–23 (2018).
- National Cancer Institute. SEER Cancer Statistics Review, 1975-2017. https://seer. cancer.gov/csr/1975\_2017/. Accessed 24 November 2020. 2020.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O. & Botstein, D. Molecular portraits of human breast tumours. *Nature*. 406, 747–752 (2000).

- Sorlie, T., Perou, C., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M., Rijn, M. v. d., Jeffrey, S., Thorsen, T., Quist, H., Matese, J., Brown, P., Botstein, D., Lonning, P. & Borresen-Dale, A. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy* of Sciences of the United States of America. 98, 10869–10874 (2001).
- Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., Livasy, C., Carey, L. A., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M. G., Sawyer, L. R., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Orrico, A. R., Dreher, D., Palazzo, J. P., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J. F., Ellis, M. J., Olopade, O. I., Bernard, P. S. & Perou, C. M. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics.* 7, 96 (2006).
- Parker, J., Mullins, M., Cheang, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron,
   C., He, X. & Hu, Z. Supervised risk predictor of breast cancer based on intrinsic subtypes.
   Journal of Clinical Oncology. 27 (2009).
- 43. Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., Ferree, S., Storhoff, J., Schaper, C. & Cuzick, J. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *Journal of Clinical Oncology.* **31**, 2783–2790 (2013).
- 44. Gnant, M., Filipits, M., Greil, R., Stoeger, H., Rudas, M., Bago-Horvath, Z., Mlineritsch, B., Kwasny, W., Knauer, M., Singer, C., Jakesz, R., Dubsky, P., Fitzal, F., Bartsch, R., Steger, G., Balic, M., Ressler, S., Cowens, J. W., Storhoff, J., Ferree, S., Schaper, C., Liu, S., Fesl, C., Nielsen, T. O., Breast, A. & Group., C. C. S. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. Annals of Oncology. 25, 339–45 (2014).
- Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., de Azambuja,
   E., Viale, G., Sotiriou, C. & Piccart, M. Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. *Journal of Clinical Oncology.* 32, 2794–2803 (2014).

- 46. Raj-Kumar, P. K., Liu, J., Hooke, J. A., Kovatich, A. J., Kvecher, L., Shriver, C. D. & Hu, H. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Scientific Reports.* 9, 7956 (2019).
- Weigelt, B., Mackay, A., A'Hern, R., Natrajan, R., Tan, D., Dowsett, M., Ashworth,
   A. & Reisfilho, J. Breast cancer molecular profiling with single sample predictors: A retrospective analysis. *The Lancet Oncology.*, 11:339–349 (2010).
- Prat, A., Parker, J., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J., He, X. & Perou, C. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research.* 12 (2010).
- Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J. & Shi, B. Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*. 5, 2929–2943 (2015).
- Holliday, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. Breast Cancer Research. 13, 215 (2011).
- Zuo, T., Zeng, H., Li, H., Liu, S., Yang, L., Xia, C., Zheng, R., Ma, F., Liu, L., Wang, N., Xuan, L. & Chen, W. The influence of stage at diagnosis and molecular subtype on breast cancer patient survival: a hospital-based multi-center study. *Chinese Journal of Cancer.* 36, 84 (2017).
- 52. Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L. & Liu, E. T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*. 100, 10393–10398 (2003).
- Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y. & Pietenpol, J. A. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*. 121, 2750–2767 (2011).

- Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A. L., Brenton, J. D., Tavare, S., Caldas, C. & Aparicio, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature.* 486, 346–352 (2012).
- Leidy, J., Khan, A. & Kandil, D. Basal-Like Breast Cancer Update on Clinicopathologic, Immunohistochemical, and Molecular Features. Archives of Pathology and Laboratory Medicine. 138, 37–43 (2014).
- Dogra, A., Mehta, A. & Doval, D. C. Are Basal-Like and Non-Basal-Like Triple-Negative Breast Cancers Really Different? *Journal of Oncology.* 2020 (2020).
- Rakha, E. A., Reis-Filho, J. S. & Ellis, I. O. Basal-like breast cancer: a critical review. Journal of Clinical Oncology. 26, 2568–2581 (2008).
- Ballinger, T., Kremer, J. & Miller, K. Triple Negative Breast Cancer-Review of Current and Emerging Therapeutic Strategies. Oncology and Hematology Review. 12, 89–94 (2016).
- Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A. L. & Botstein, D. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America.* 100 (2003).
- Potemski, P., Kusinska, R., Watala, C., Pluciennik, E., Bednarek, A. K. & Kordek, R. Prognostic relevance of basal cytokeratin expression in operable breast cancer. *Oncology.* 69, 478–485 (2005).
- Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca,
   G., Troester, M. A., Tse, C. K., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C.,
   Nielsen, T. O., Moorman, P. G., Earp, H. S. & Millikan, R. C. Race, breast cancer

subtypes, and survival in the Carolina Breast Cancer Study. *Journal of the American Medical Association.* **295**, 2492–2502. (2006).

- Foulkes, W. D., Brunet, J. S., Stefansson, I. M., Straume, O., Chappuis, P. O., Begin, L. R., Hamel, N., Goffin, J. R., Wong, N., Trudel, M., Kapusta, L., Porter, P. & Akslen, L. A. The prognostic implication of the basal-like phenotype of BRCA1-related breast cancer. *Cancer Research.* 64, 830–835 (2004).
- Banerjee, S., Reis-Filho, J. S., Ashley, S., Steele, D., Ashworth, A., Lakhani, S. R. & Smith, I. E. Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *Journal of Clinical Pathology.* 59, 729–735 (2006).
- Rakha, E. A., El-Rehim, D. A., Paish, C., Green, A. R., Lee, A. H., Robertson, J. F., Blamey, R. W., Macmillan, D. & Ellis, I. O. Basal phenotype identifies a poor prognostic subgroup of breast cancer of clinical importance. *European Journal of Cancer.* 42, 3149– 3156 (2006).
- Toft, D. J. & Cryns, V. L. Minireview: Basal-like breast cancer: from molecular profiles to targeted therapies. *Molecular Endocrinology.* 25, 199–211. (2011).
- Allott, E. H., Geradts, J., Cohen, S. M., Khoury, T., Zirpoli, G. R., Bshara, W., Davis, W., Omilian, A., Nair, P., Ondracek, R. P., Cheng, T. D., Miller, C. R., Hwang, H., Thorne, L. B., OConnor, S., Bethea, T. N., Bell, M. E., Hu, Z., Li, Y., Kirk, E. L., Sun, X., Ruiz-Narvaez, E. A., Perou, C. M., Palmer, J. R., Olshan, A. F., Ambrosone, C. B. & Troester, M. A. Frequency of breast cancer subtypes among African American women in the AMBER consortium. *Breast Cancer Research.* 20, 12 (2018).
- 67. Hicks, D. G., Short, S. M., Prescott, N. L., Tarr, S. M., Coleman, K. A., Yoder, B. J., Crowe, J. P., Choueiri, T. K., Dawson, A. E., Budd, G. T., Tubbs, R. R., Casey, G. & Weil, R. J. Breast cancers with brain metastases are more likely to be estrogen receptor negative, express the basal cytokeratin CK5/6, and overexpress HER2 or EGFR. *The American Journal of Surgical Pathology.* **30**, 1097–1104 (2006).

- Fulford, L. G., Reis-Filho, J. S., Ryder, K., Jones, C., Gillett, C. E., Hanby, A., Easton,
   D. & Lakhani, S. R. Basal-like grade III invasive ductal carcinoma of the breast: patterns of metastasis and long-term survival. *Breast Cancer Research.* 9, R4 (2007).
- Walter, O., Prasad, M., Lu, S., Quinlan, R. M., Edmiston, K. L. & Khan, A. IMP3 is a novel biomarker for triple negative invasive mammary carcinoma associated with a more aggressive phenotype. *Human Pathology.* 40, 1528–1533 (2009).
- Rakha, E. A., Elsheikh, S. E., Aleskandarany, M. A., Habashi, H. O., Green, A. R., Powe, D. G., El-Sayed, M. E., Benhasouna, A., Brunet, J. S., Akslen, L. A., Evans, A. J., Blamey, R., Reis-Filho, J. S., Foulkes, W. D. & Ellis, I. Triple-negative breast cancer: distinguishing between basal and nonbasal subtypes. *Clinical Cancer Research*. 15, 2302–2310 (2009).
- Nielsen, T. O., Hsu, F. D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., Dressler, L., Akslen, L. A., Ragaz, J., Gown, A. M., Gilks, C. B., van de Rijn, M. & Perou, C. M. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clinical Cancer Research*. 10, 5367–5374. (2004).
- Den Hollander, P., Savage, M. I. & Brown, P. H. Targeted therapy for breast cancer prevention. *Frontiers in Oncology.* 3, 250 (2013).
- Davies, C., Godwin, J., Gray, R., Clarke, M., Cutter, D., Darby, S., McGale, P., Pan, H. C., Taylor, C., Wang, Y. C., Dowsett, M., Ingle, J. & Peto, R. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patientlevel meta-analysis of randomised trials. *The Lancet.* 378, 771–784 (2011).
- Cuzick, J., Powles, T., Veronesi, U., Forbes, J., Edwards, R., Ashley, S. & Boyle, P. Overview of the main outcomes in breast-cancer prevention trials. *The Lancet.* 361, 296–300 (2003).
- 75. Cummings, S. R., Eckert, S., Krueger, K. A., Grady, D., Powles, T. J., Cauley, J. A., Norton, L., Nickelsen, T., Bjarnason, N. H., Morrow, M., Lippman, M. E., Black, D., Glusman, J. E., Costa, A. & Jordan, V. C. The effect of raloxifene on risk of breast

cancer in postmenopausal women: results from the MORE randomized trial. Multiple Outcomes of Raloxifene Evaluation. *Journal of the American Medical Association.* **281**, 2189–97. (1999).

- 76. Martino, S., Cauley, J. A., Barrett-Connor, E., Powles, T. J., Mershon, J., Disch, D., Secrest, R. J. & Cummings, S. R. Continuing outcomes relevant to Evista: breast cancer incidence in postmenopausal osteoporotic women in a randomized trial of raloxifene. *Journal of the National Cancer Institute.* **96**, 1751–1761 (2004).
- Barrett-Connor, E., Mosca, L., Collins, P., Geiger, M. J., Grady, D., Kornitzer, M., Mc-Nabb, M. A. & Wenger, N. K. Effects of raloxifene on cardiovascular events and breast cancer in postmenopausal women. *The New England Journal of Medicine*. 355, 125–137 (2006).
- Cummings, S. R., Ensrud, K., Delmas, P. D., LaCroix, A. Z., Vukicevic, S., Reid, D. M., Goldstein, S., Sriram, U., Lee, A., Thompson, J., Armstrong, R. A., Thompson, D. D., Powles, T., Zanchetta, J., Kendler, D., Neven, P. & Eastell, R. Lasofoxifene in postmenopausal women with osteoporosis. *The New England Journal of Medicine.* 362, 686– 96 (2010).
- Slamon, D., Eiermann, W., Robert, N., Pienkowski, T., Martin, M., Press, M., Mackey, J., Glaspy, J., Chan, A., Pawlicki, M., Pinter, T., Valero, V., Liu, M. C., Sauter, G., Minckwitz, G., Visco, F., Bee, V., Buyse, M., Bendahmane, B., Tabah-Fisch, I., Lindsay, M. A., Riva, A. & Crown, J. Adjuvant trastuzumab in HER2-positive breast cancer. *The New England Journal of Medicine.* 365, 1273–1283 (2011).
- Cuello, M., Ettenberg, S. A., Clark, A. S., Keane, M. M., Posner, R. H., Nau, M. M., Dennis, P. A. & Lipkowitz, S. Down-regulation of the erbB-2 receptor by trastuzumab (herceptin) enhances tumor necrosis factor-related apoptosis-inducing ligand-mediated apoptosis in breast and ovarian cancer cell lines that overexpress erbB-2. *Cancer Research.* 61, 4892–900 (2001).
- Junttila, T. T., Akita, R. W., Parsons, K., Fields, C., Lewis Phillips, G. D., Friedman,
   L. S., Sampath, D. & Sliwkowski, M. X. Ligand-independent HER2/HER3/PI3K complex

is disrupted by trastuzumab and is effectively inhibited by the PI3K inhibitor GDC-0941. Cancer Cell. 15, 429–440 (2009).

- Yakes, F. M., Chinratanalab, W., Ritter, C. A., King, W., Seelig, S. & Arteaga, C. L. Herceptin-induced inhibition of phosphatidylinositol-3 kinase and Akt Is required for antibody-mediated effects on p27, cyclin D1, and antitumor action. *Cancer Research.* 62, 4132–4141 (2002).
- Geyer, C. E., Forster, J., Lindquist, D., Chan, S., Romieu, C. G., Pienkowski, T., Jagiello-Gruszfeld, A., Crown, J., Chan, A., Kaufman, B., Skarlos, D., Campone, M., Davidson, N., Berger, M., Oliva, C., Rubin, S. D., Stein, S. & Cameron, D. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *The New England Journal of Medicine*. 355, 2733–2743 (2006).
- Verma, S., Miles, D., Gianni, L., Krop, I. E., Welslau, M., Baselga, J., Pegram, M., Oh, D. Y., Dieras, V., Guardino, E., Fang, L., Lu, M. W., Olsen, S. & Blackwell, K. Trastuzumab emtansine for HER2-positive advanced breast cancer. *The New England Journal of Medicine.* 367, 1783–1791 (2012).
- Swain, S. M., Baselga, J., Kim, S. B., Ro, J., Semiglazov, V., Campone, M., Ciruelos, E., Ferrero, J. M., Schneeweiss, A., Heeson, S., Clark, E., Ross, G., Benyunes, M. C. & Cortes, J. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *The New England Journal of Medicine*. **372**, 724–734 (2015).
- 86. Brufsky, A. M., Hurvitz, S., Perez, E., Swamy, R., Valero, V., O'Neill, V. & Rugo, H. S. RIBBON-2: a randomized, double-blind, placebo-controlled, phase III trial evaluating the efficacy and safety of bevacizumab in combination with chemotherapy for second-line treatment of human epidermal growth factor receptor 2-negative metastatic breast cancer. *Journal of Clinical Oncology.* 29, 4286–4293 (2011).
- 87. Robert, N. J., Dieras, V., Glaspy, J., Brufsky, A. M., Bondarenko, I., Lipatov, O. N., Perez,
  E. A., Yardley, D. A., Chan, S. Y., Zhou, X., Phan, S. C. & OShaughnessy, J. RIBBON1: randomized, double-blind, placebo-controlled, phase III trial of chemotherapy with or
  without bevacizumab for first-line treatment of human epidermal growth factor receptor

2-negative, locally recurrent or metastatic breast cancer. Journal of Clinical Oncology.29, 1252–1260 (2011).

- 88. Mackey, J. R., Ramos-Vazquez, M., Lipatov, O., McCarthy, N., Krasnozhon, D., Semiglazov, V., Manikhas, A., Gelmon, K. A., Konecny, G. E., Webster, M., Hegg, R., Verma, S., Gorbunova, V., Abi Gerges, D., Thireau, F., Fung, H., Simms, L., Buyse, M., Ibrahim, A. & Martin, M. Primary results of ROSE TRIO-12, a randomized placebo-controlled phase III trial evaluating the addition of ramucirumab to first-line docetaxel chemotherapy in metastatic breast cancer. *Journal of Clinical Oncology.* 33, 141–148 (2015).
- Bergh, J., Bondarenko, I. M., Lichinitser, M. R., Liljegren, A., Greil, R., Voytko, N. L., Makhson, A. N., Cortes, J., Lortholary, A., Bischoff, J., Chan, A., Delaloge, S., Huang, X., Kern, K. A. & Giorgetti, C. First-line treatment of advanced breast cancer with sunitinib in combination with docetaxel versus docetaxel alone: results of a prospective, randomized phase III study. *Journal of Clinical Oncology.* **30**, 921–929 (2012).
- 90. Crown, J. P., Dieras, V., Staroslawska, E., Yardley, D. A., Bachelot, T., Davidson, N., Wildiers, H., Fasching, P. A., Capitain, O., Ramos, M., Greil, R., Cognetti, F., Fountzilas, G., Blasinska-Morawiec, M., Liedtke, C., Kreienberg, R., Miller, W. H., Jr Tassell, V., Huang, X., Paolini, J., Kern, K. A. & Romieu, G. Phase III trial of sunitinib in combination with capecitabine versus capecitabine monotherapy for the treatment of patients with pretreated metastatic breast cancer. *Journal of Clinical Oncology.* **31**, 2870–2878 (2013).
- 91. Baselga, J., Costa, F., Gomez, H., Hudis, C. A., Rapoport, B., Roche, H., Schwartzberg, L. S., Petrenciuc, O., Shan, M. & Gradishar, W. J. A phase 3 trial comparing capecitabinE in combination with SorafenIb or placebo for treatment of locally advanced or metastatIc HER2-Negative breast Cancer the RESILIENCE study: study protocol for a randomized controlled trial. Trials. 14 (2013).
- 92. Baselga, J., Gomez, P., Greil, R., Braga, S., Climent, M. A., Wardley, A. M., Kaufman, B., Stemmer, S. M., Pego, A., Chan, A., Goeminne, J. C., Graas, M. P., Kennedy, M. J., Ciruelos Gil, E. M., Schneeweiss, A., Zubel, A., Groos, J., Melezinkova, H. & Awada, A. Randomized phase II study of the anti-epidermal growth factor receptor monoclonal an-

tibody cetuximab with cisplatin versus cisplatin alone in patients with metastatic triplenegative breast cancer. *Journal of Clinical Oncology.* **31**, 2586–2592 (2013).

- 93. Carey, L. A., Rugo, H. S., Marcom, P. K., Mayer, E. L., Esteva, F. J., Ma, C. X., Liu, M. C., Storniolo, A. M., Rimawi, M. F., Forero-Torres, A., Wolff, A. C., Hobday, T. J., Ivanova, A., Chiu, W. K., Ferraro, M., Burrows, E., Bernard, P. S., Hoadley, K. A., Perou, C. M. & Winer, E. P. 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. (2012).
- 94. Baselga, J., Albanell, J., Ruiz, A., Lluch, A., Gascon, P., Guillem, V., Gonzalez, S., Sauleda, S., Marimon, I., Tabernero, J. M., Koehler, M. T. & Rojo, F. Phase II and tumor pharmacodynamic study of gefitinib in patients with advanced breast cancer. *Journal of Clinical Oncology.* 23, 5323–5333 (2005).
- Dickler, M. N., Rugo, H. S., Eberle, C. A., Brogi, E., Caravelli, J. F., Panageas, K. S., Boyd, J., Yeh, B., Lake, D. E., Dang, C. T., Gilewski, T. A., Bromberg, J. F., Seidman, A. D., D'Andrea, G. M., Moasser, M. M., Melisko, M., Park, J. W., Dancey, J., Norton, L. & Hudis, C. A. A phase II trial of erlotinib in combination with bevacizumab in patients with metastatic breast cancer. *Clinical Cancer Research.* 14, 7878–7883 (2008).
- 96. Tutt, A., Robson, M., Garber, J. E., Domchek, S. M., Audeh, M. W., Weitzel, J. N., Friedlander, M., Arun, B., Loman, N., Schmutzler, R. K., Wardley, A., Mitchell, G., Earl, H., Wickens, M. & Carmichael, J. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. *The Lancet.* **376**, 235–244 (2010).
- Evans, R. M. & Mangelsdorf, D. J. Nuclear Receptors, RXR, and the Big Bang. Cell. 157, 255–266 (2014).
- 98. Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. & Evans, R. M. The nuclear receptor superfamily: the second decade. *Cell.* 83, 835–839 (1995).

- Nuclear Receptors Nomenclature Committee. A unified nomenclature system for the nuclear receptor superfamily. *Cell.* 97, 161–163 (1999).
- Owen, G. I. & Zelent, A. Origins and evolutionary diversification of the nuclear receptor superfamily. *Cellular and Molecular Life Sciences.* 57, 809–827 (2000).
- Germain, P., Staels, B., Dacquet, C., Spedding, M. & Laudet, V. Overview of nomenclature of nuclear receptors. *Pharmacological Reviews.* 58, 685–704 (2006).
- Doan, T. B., Graham, J. D. & Clarke, C. L. Emerging functional roles of nuclear receptors in breast cancer. *Journal of Molecular Endocrinology.* 58, R169–R190 (2017).
- Weikum, E. R., Liu, X. & Ortlund, E. A. The nuclear receptor superfamily: A structural perspective. *Protein Science.* 27, 1876–1892 (2018).
- 104. Gee, A. C. & Katzenellenbogen, J. A. Probing conformational changes in the estrogen receptor: evidence for a partially unfolded intermediate facilitating ligand binding and release. *Molecular Endocrinology.* 15(, 421–428. (2001).
- 105. Johnson, B. A., Wilson, E. M., Li, Y., Moller, D. E., Smith, R. G. & Zhou, G. Ligandinduced stabilization of PPARgamma monitored by NMR spectroscopy: implications for nuclear receptor activation. *Journal of Molecular Biology.* 298, 187–194 (2000).
- 106. Kallenberger, B. C., Love, J. D., Chatterjee, V. K. & Schwabe, J. W. A dynamic mechanism of nuclear receptor activation and its perturbation in a human disease. *Nature Structural and Molecular Biology.* 10, 136–140 (2003).
- Darimont, B. D., Wagner, R. L., Apriletti, J. W., Stallcup, M. R., Kushner, P. J., Baxter,
   J. D., Fletterick, R. J. & Yamamoto, K. R. Structure and specificity of nuclear receptorcoactivator interactions. *Genes and Development.* 12, 3343–3356 (1998).
- Millard, C. J., Watson, P. J., Fairall, L. & Schwabe, J. W. An evolving understanding of nuclear receptor coregulator proteins. *Journal of Molecular Endocrinology.* 51, T23–36 (2013).
- McKenna, N. J. & O'Malley, B. W. Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell.* 108, 465–474 (2002).

- Halachmi, S., Marden, E., Martin, G., MacKay, H., Abbondanza, C. & Brown, M. Estrogen receptor-associated proteins: possible mediators of hormone-induced transcription. *Science.* 264, 1455–1458 (1994).
- 111. Baniahmad, A., Leng, X., Burris, T. P., Tsai, S. Y., Tsai, M. J. & O'Malley, B. W. The tau 4 activation domain of the thyroid hormone receptor is required for release of a putative corepressor(s) necessary for transcriptional silencing. *Molecular and Cellular Biology.* 15, 76–86 (1995).
- Cavailles, V., Dauvois, S., L'Horset, F., Lopez, G., Hoare, S., Kushner, P. J. & Parker, M. G. Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. *The EMBO Journal.* 14, 3741–3751 (1995).
- Lonard, D. M. & O'Malley, B. W. Nuclear receptor coregulators: modulators of pathology and therapeutic targets. *Nature Reviews Endocrinology*. 8, 598–604 (2012).
- 114. R, R. D., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T. R., Ghosh, D. & Chinnaiyan, A. M. Mining for regulatory programs in the cancer transcriptome. *Nature Genetics.* 37, 579–583. (2005).
- Graham, J. D., Bain, D. L., Richer, J. K., Jackson, T. A., Tung, L. & Horwitz, K. B. Nuclear receptor conformation, coregulators, and tamoxifen-resistant breast cancer. *Steroids*. 65, 579–584 (2000).
- 116. Keeton, E. K. & Brown, M. Coregulator expression and breast cancer: improving the predictive power of estrogen receptor alpha. *Clinical Cancer Research.* 9, 1229–1230 (2003).
- 117. Girault, I., Lerebours, F., Amarir, S., Tozlu, S., Tubiana-Hulin, M., Lidereau, R. & Bieche,
  I. Expression analysis of estrogen receptor alpha coregulators in breast carcinoma: evidence that NCOR1 expression is predictive of the response to tamoxifen. *Clinical Cancer Research.* 9, 1259–1266 (2003).
- 118. Doan, T. B., Eriksson, N. A., Graham, D., Funder, J. W., Simpson, E. R., Kuczek, E. S., Clyne, C., Leedman, P. J., Tilley, W. D., Fuller, P. J., Muscat, G. E. & Clarke, C. L. Breast cancer prognosis predicted by nuclear receptor-coregulator networks. *Molecular Oncology.* 8, 998–1013 (2014).

- Cordera, F. & Jordan, V. C. Steroid receptors and their role in the biology and control of breast cancer growth. *Seminars in Oncology.* 33, 631–641 (2006).
- 120. Barton, V. N., D'Amato, N. C., Gordon, M. A., Christenson, J. L., Elias, A. & Richer, J. K. Androgen Receptor Biology in Triple Negative Breast Cancer: a Case for Classification as AR+ or Quadruple Negative Disease. *Hormones and Cancer.* 6, 206–213 (2015).
- 121. Krishnan, A. V., Swami, S. & Feldman, D. Vitamin D and breast cancer: inhibition of estrogen synthesis and signaling. *The Journal of Steroid Biochemistry and Molecular Biology.* **121**, 343–348 (2010).
- 122. Mehta, R. G., Peng, X., Alimirah, F., Murillo, G. & Mehta, R. Vitamin D and breast cancer: emerging concepts. *Cancer Letters.* 334, 95–100 (2013).
- 123. Vilasco, M., Communal, L., Mourra, N., Courtin, A., Forgez, P. & Gompel, A. Glucocorticoid receptor and breast cancer. Breast Cancer Research and Treatment. 130, 1–10 (2011).
- 124. Abduljabbar, R., Negm, O. H., Lai, C. F., Jerjees, D. A., Al-Kaabi, M., Hamed, M. R., Tighe, P. J., Buluwela, L., Mukherjee, A., Green, A. R., Ali, S., Rakha, E. A. & Ellis, I. O. Clinical and biological significance of glucocorticoid receptor (GR) expression in breast cancer. *Breast Cancer Research and Treatment.* 150, 335–346 (2015).
- 125. Liu, Y., Lee, M. O., Wang, H. G., Li, Y., Hashimoto, Y., Klaus, M., Reed, J. C. & Zhang, X. Retinoic acid receptor beta mediates the growth-inhibitory effect of retinoic acid by promoting apoptosis in human breast cancer cells. *Molecular and Cellular Biology.* 16, 1138–1149 (1996).
- 126. Yang, Q., Sakurai, T. & Kakudo, K. Retinoid, retinoic acid receptor beta and breast cancer. Breast Cancer Research and Treatment. 76, 167–73. (2002).
- Flamini, M. I., Gauna, G. V., Sottile, M. L., Nadin, B. S., Sanchez, A. M. & Vargas-Roig,
   L. M. Retinoic acid reduces migration of human breast cancer cells: role of retinoic acid receptor beta. *Journal of Cellular and Molecular Medicine*. 18, 1113–1123 (2014).

- 128. Akerstedt, T., Knutsson, A., Narusyte, J., Svedberg, P., Kecklund, G. & Alexanderson, K. Night work and breast cancer in women: a Swedish cohort study. *BMJ Open.* 5, e008127 (2015).
- 129. Akashi, M. & Takumi, T. The orphan nuclear receptor RORalpha regulates circadian transcription of the mammalian core-clock Bmal1. Nature Structural and Molecular Biology. 12, 441–448 (2005).
- 130. Preitner, N., Damiola, F., Lopez-Molina, L., Zakany, J., Duboule, D., Albrecht, U. & Schibler, U. The orphan nuclear receptor REV-ERBalpha controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell.* **110**, 251–260 (2002).
- 131. Conway-Campbell, B. L., Sarabdjitsingh, R. A., McKenna, M. A., Pooley, J. R., Kershaw, Y. M., Meijer, O. C., de Kloet, E. R. & Lightman, S. L. Glucocorticoid ultradian rhythmicity directs cyclical gene pulsing of the clock gene period 1 in rat hippocampus. *Journal of Neuroendocrinology.* 22, 1093–1100 (2010).
- 132. Dufour, C. R., Levasseur, M. P., Pham, N. H., Eichner, L. J., Wilson, B. J., Charest-Marcotte, A., Duguay, D., Poirier-Heon, J. F., Cermakian, N. & Giguere, V. Genomic convergence among ERRα, PROX1, and BMAL1 in the control of metabolic clock outputs. *PLoS Genet.* 7, e1002143 (2011).
- 133. Cho, H., Zhao, X., Hatori, M., Yu, R. T., Barish, G. D., Lam, M. T., Chong, L. W., DiTacchio, L., Atkins, A. R., Glass, C. K., Liddle, C., Auwerx, J., Downes, M., Panda, S. & Evans, R. M. Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. Nature. 485, 123–127 (2012).
- Gerhart-Hines, Z. & Lazar, M. A. Rev-erbα and the circadian transcriptional regulation of metabolism. *Diabetes, Obesity and Metabolism.* 17 Suppl 1, 12–16 (2015).
- Giguere, V. Transcriptional control of energy homeostasis by the estrogen-related receptors. *Endocrine Reviews.* 29, 677–696 (2008).
- Deblois, G. & Giguere, V. Oestrogen-related receptors in breast cancer: control of cellular metabolism and beyond. *Nature Reviews Cancer.* 13, 27–36 (2013).

- Michalik, L., Desvergne, B. & Wahli, W. Peroxisome-proliferator-activated receptors and cancers: complex stories. *Nature Reviews Cancer.* 4, 61–70 (2004).
- Ward, P. S. & Thompson, C. B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell.* 21, 297–308 (2012).
- 139. De Mei, C., Ercolani, L., Parodi, C., Veronesi, M., Lo Vecchio, C., Bottegoni, G., Torrente, E., Scarpelli, R., Marotta, R., Ruffili, R., Mattioli, M., Reggiani, A., Wade, M. & Grimaldi, B. Dual inhibition of REV-ERBβ and autophagy as a novel pharmacological approach to induce cytotoxicity in cancer cells. *Oncogene.* **34**, 2597–2608 (2015).
- Cai, Q., Lin, T., Kamarajugadda, S. & Lu, J. Regulation of glycolysis and the Warburg effect by estrogen-related receptors. *Oncogene.* 32, 2079–2086. (2013).
- Sakharkar, M. K., Shashni, B., Sharma, K., Dhillon, S. K., Ranjekar, P. R. & Sakharkar,
  K. R. Therapeutic implications of targeting energy metabolism in breast cancer. *PPAR Research.* 2013, 109285 (2013).
- Avena, P., Anselmo, W., Whitaker-Menezes, D., Wang, C., Pestell, R. G., S, L. R., Hulit, J., Casaburi, I., Ando, S., Martinez-Outschoorn, U. E., Lisanti, M. P. & Sotgia, F. Compartment-specific activation of PPARγ governs breast cancer tumor growth, via metabolic reprogramming and symbiosis. *Cell Cycle.* 12, 1360–1370 (2013).
- 143. Li, H., Tu, Z., An, L., Qian, Z., Achilefu, S. & Gu, Y. Inhibitory effects of ERβ on proliferation, invasion, and tumor formation of MCF-7 breast cancer cells-prognostication for the use of ERβ-selective therapy. *Pharmaceutical Biology.* **50**, 839–849 (2012).
- 144. McGowan, E. M., Saad, S., Bendall, L. J., Bradstock, K. F. & Clarke, C. L. Effect of progesterone receptor a predominance on breast cancer cell migration into bone marrow fibroblasts. *Breast Cancer Research and Treatment.* 83, 211–220. (2004).
- 145. Carnevale, R. P., Proietti, C. J., Salatino, M., Urtreger, A., Peluffo, G., Edwards, D. P., Boonyaratanakornkit, V., Charreau, E. H., Bal de Kier Joffe, E., Schillaci, R. & Elizalde, P. V. Progestin effects on breast cancer cell proliferation, proteases activation, and in vivo development of metastatic phenotype all depend on progesterone receptor capacity to activate cytoplasmic signaling pathways. *Molecular Endocrinology.* 21, 1335–1358 (2007).

- Misawa, A. & Inoue, S. Estrogen-Related Receptors in Breast Cancer and Prostate Cancer. Frontiers in Endocrinology. 6, 83 (2015).
- 147. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. Journal of the American Medical Association. 323, 844–853 (2020).
- 148. Seyhan, A. A. Lost in translation: the valley of death across preclinical and clinical divide
   identification of problems and overcoming obstacles. *Translational Medicine Communi*cations. 4 (2019).
- 149. Williams, K., Bilsland, E., Sparkes, A., Aubrey, W., Young, M., Soldatova, L. N., De Grave, K., Ramon, J., de Clare, M., Sirawaraporn, W., Oliver, S. G. & King, R. D. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society Interface.* 12, 20141289 (2015).
- Pina, A. S., Hussain, A. & Roque, A. C. An historical overview of drug discovery. *Methods in Molecular Biology.* 572, 3–12 (2009).
- 151. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology*.
  18, 83 (2017).
- He, K. Y., Ge, D. & He, M. M. Big Data Analytics for Genomic Medicine. International Journal of Molecular Sciences. 18, 412 (2017).
- Trosset, J. Y. & Vodovar, N. Structure-based target druggability assessment. Methods in Molecular Biology. 986, 141–164 (2013).
- Trosset, J. Y. & Cave, C. In Silico Target Druggability Assessment: From Structural to Systemic Approaches. *Methods in Molecular Biology.* 1953, 63–88 (2019).
- 155. Steinmetz, K. L. & Spack, E. G. The basics of preclinical drug development for neurodegenerative disease indications. *BMC Neurology*. **9 Suppl 1**, S2 (2009).
- 156. Romano, J. D. & Tatonetti, N. P. Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives. *Frontiers in Genetics.* 10 (2019).

- 157. Tanoli, Z., Alam, Z., Vaha-Koskela, M., Ravikumar, B., Malyutina, A., Jaiswal, A., Tang, J., Wennerberg, K. & Aittokallio, T. Drug Target Commons 2.0: a community platform for systematic analysis of drug-target interaction profiles. *Database(Oxford).*, 1–13 (2018).
- Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., Pierleoni, A., Pignatelli, M., Platt, T., Rowland, F., Wankar, P., Bento, A. P., Burdett, T., Fabregat, A., Forbes, S., Gaulton, A., Gonzalez, C. Y., Hermjakob, H., Hersey, A., Jupe, S., Kafkas, S., Keays, M., Leroy, C., Lopez, F. J., Magarinos, M. P., Malone, J., McEntyre, J., Munoz-Pomer Fuentes, A., O'Donovan, C., Papatheodorou, I., Parkinson, H., Palka, B., Paschall, J., Petryszak, R., Pratanwanich, N., Sarntivijal, S., Saunders, G., Sidiropoulos, K., Smith, T., Sondka, Z., Stegle, O., Tang, Y. A., Turner, E., Vaughan, B., Vrousgou, O., Watkins, X., Martin, M. J., Sanseau, P., Vamathevan, J., Birney, E., Barrett, J. & Dunham, I. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research.* 45, 985–994 (2017).
- Paananen, J. & Fortino, V. An omics perspective on drug target discovery platforms. Briefings in Bioinformatics. bbz122, 1–17 (2019).
- 160. Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I. & Overington, J. P. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery.* 16, 19–34 (2017).
- Gashaw, I., Ellinghaus, P., Sommer, A. & Asadullah, K. What makes a good drug target? Drug Discov Today. 16, 1037–43 (2011).
- Bakheet, T. M. & Doig, A. J. Properties and identification of human protein drug targets. Bioinformatics. 25, 451–457 (2009).
- Imming, P., Sinning, C. & Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nature Reviews Drug Discovery.* 5, 821–834 (2006).
- 164. Smith, C. Drug target identification: a question of biology. Nature. 428, 225–31 (2004).
- 165. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. Molecular classification

of cancer: class discovery and class prediction by gene expression monitoring. *Science*. **286**, 531–537 (1999).

- 166. Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J. J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 403, 503–511 (2000).
- 167. Therese, S., Robert, T., Joel, P., Trevor, H., Marron, J., Andrew, N., Shibing, D., Hilde, J., Robert, P. & Stephanie, G. Stephanie G. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* of the United States of America. 100, 8418–8423 (2003).
- Therese, S. Molecular portraits of breast cancer: Tumour subtypes as distinct disease entities. *European Journal of Cancer.* 40, 2667–2675 (2004).
- 169. Kumar, N., Zhao, D., Bhaumik, D., Sethi, A. & Gann, P. Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes. BMC Cancer. 19, 215 (2019).
- Datta, S. & Datta, S. Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics. 7 (2006).
- 171. Xu, R., Damelin, S., Nadler, B. & Wunsch, D. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. Artificial Intelligence in Medicine. 48, 91–98 (2010).
- 172. Vidman, V., Kallberg, D. & Ryden, P. Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLoS One.* 14, e0219102 (2019).
- Sokal, R. & Michener, C. A statistical method for evaluating systematic relationships.
   The University of Kansas Science Bulletin. 38, 1409–1438 (19158).

- McQuitty, L. Hierarchical linkage analysis for the isolation of types. Educational and Psychological Measurement. 20, 55–67 (1960).
- Sneath, P. & Sokal, R. R. Numerical taxonomy: theprinciples and practice of numerical classification. San Francisco: Freeman., 359 (1963).
- 176. Dhaeseleer<sub>2</sub>005. How does gene expression clustering work? Nature Biotechnology. 23, 1499–1502 (2005).
- 177. Jiang, D., Tang, C. & Zhang, A. Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering. 16, 1370–1386 (2004).
- Armstrong, R. A. Should Pearson's correlation coefficient be avoided? Ophthalmic and Physiological Optics. 39, 316–327 (2019).
- 179. Ma, G., Ma, C. & Wu, J. Data Clustering: Theory, Algorithms, and Applications (eds Gan, G., Ma, C. & Wu, J.) (Society for Inductrial and Applied Mathematics, pu, 2007).
- Fraley, C. & Adrian, E. R. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 97, 611–631 (2002).
- Evans, K., Love, T. & Thurston, S. W. Outlier Identification in Model-Based Cluster Analysis. *The Journal of Classification.* **32**, 63–84 (2015).
- Heller, K. A. & Ghahramani, Z. Bayesian Hierarchical Clustering. in Twenty-second International Conference on Machine Learning (2005).
- 183. Cooke, E., Savage, R., Kirk, P., Darkins, R. & Wild, D. Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics.* 12, 399 (2011).
- 184. Savage, R., Heller, K., Xu, Y., Ghahramani, Z., Truman, W., Grant, M., Denby, K. & Wild, D. R/BHC: fast Bayesian hierarchical clustering for microarray data. BMC Bioinformatics. 10, 242 (2009).
- 185. De, S. M., Costa, I., De, A. D., Ludermir, T. & Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics.* 9, 497 (2008).

- 186. Sirinukunwattana, K., Savage, R. S., Bari, M. F., Snead, D. R. & Rajpoot, N. M. Bayesian hierarchical clustering for studying cancer gene expression data with unknown statistics. *PLoS One.* 8, e75748 (2013).
- 187. Bild, A. H., Potti, A. & Nevins, J. R. Linking oncogenic pathways with therapeutic opportunities. *Nature Reviews Cancer.* 6, 735–41 (2006).
- 188. Itadani, H., Mizuarai, S. & Kotani, H. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. *Current Genomics.* 9, 349–360 (2008).
- Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V. & Lankelma, J. Cancer: a systems biology disease. *Biosystems.* 83, 81–90 (2006).
- Ma'ayan, A. Introduction to Network Analysis in Systems Biology. Science Signaling. 4 (2011).
- Blasius, B. & Brockmann, D. Frontiers in network science: advances and applications. The European Physical Journal. 84, 491–492 (2011).
- 192. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. **302**, 249–255 (2003).
- 193. Applied Statistics for Network Biology: Methods in Systems Biology (Wiley-Blackwell, 2011).
- Roy, S., Bhattacharyya, D. K. & Kalita, J. K. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics*. 15 (2014).
- 195. Dam, S., Vosa, U., Graaf, A., Franke, L. & Magalhaes, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*. 19, 575–592 (2018).
- 196. Oh, E.-Y., Christensen, S. M., Ghanta, S., Jeong, J. C., Bucur, O., Glass, B., Montaser-Kouhsari, L., Knoblauch, N. W., Bertos, N., Saleh, S. M., Haibe-Kains, B., Park, M.

& Beck, A. H. Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biology.* **16** (2015).

- 197. Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N. & Liang, H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications.* 5 (2014).
- Ahmadi, A. A. & Qian, X. Tumor stratification by a novel graph-regularized bi-clique finding algorithm. *Computational Biology and Chemistry.* 57, 3–11 (2015).
- Koboldt, D., Fulton, R. & McLellan, M. Comprehensive molecular portraits of human breast tumours. *Nature*. 490, 61–70 (2012).
- 200. Pereira, B., Chin, S., Rueda, O., Vollan, H., Provenzano, E., Bardwell, H., Pugh, M., Jones, L., Russell, R., Sammut, S., Tsui, D., Liu, B., Dawson, S., Abraham, J., Northen, H., Peden, J., Mukherjee, A., Turashvili, G., Green, A., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D., Ellis, I., Purushotham, A., Pinder, S., Borresen-Dale, A., Earl, H., Pharoah, P., Ross, M., Aparicio, S. & Caldas, C. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications.* 10, 11479 (2016).
- 201. Rueda, O., Sammut, S., Seoane, J., Chin, S., Caswell-Jin, J., Callari, M., Batra, R., Pereira, B., Bruna, A., Ali, H., Provenzano, E., Liu, B., Parisien, M., Gillett, C., McKinney, S., Green, A., Murphy, L., Purushotham, A., Ellis, I., Pharoah, P., Rueda, C., Aparicio, S., Caldas, C. & Curtis, C. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature.* 567, 399–404 (2019).
- 202. Opgen-Rhein, R. & Strimmer, K. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT.* 4, 53–65 (2006).
- 203. Schafer, J. & Strimmer, K. A shrinkage approach to large-scale covariance estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology.* 4, 32 (2005).
- 204. Wang, Y., Kang, J., Kemmer, P. & Gu, o. Y. An Efficient and Reliable Statistical Method for Estimating Functional Connectivity in Large Scale Brain Networks Using Partial Correlation. *Frontiers in Neuroscience*. **31**, 123 (2016).
- 205. Johansson, A., Loset, M., Mundal, S. B., Johnson, M. P., Freed, K. A., Fenstad, M. H., Moses, E. K., Austgulen, R. & Blangero, J. Partial correlation network analyses to detect altered gene interactions in human disease: using preeclampsia as a model. *Human Genetics.* **129**, 25–34 (2011).
- 206. De la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics.* 20, 3565–74 (2004).
- 207. Han, L. & Zhu, J. Using matrix of thresholding partial correlation coefficients to infer regulatory network. *Biosystems.* **91** (2008).
- 208. Fujita, A., Sato, J. R., Garay-Malpartida, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C. & Ferreira, C. E. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology.* **30**, 39 (2007).
- Schafer, J. & Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics.* 21(6): 754–764 (2005).
- Prill, R. J., Iglesias, P. A. & Levchenko, A. Dynamic properties of network motifs contribute to biological network organization. *PLOS Biology.* 3, e343 (2005).
- Davis, J. A. & Leinhardt, S. The Structure of Positive Interpersonal Relations in Small Groups. (ed J, I. B.) 218-251 (Boston: Houghton Mifflin, 1972).
- Polyak, K. Heterogeneity in breast cancer. Journal of Clinical Investigation. 121, 3786– 3788 (2011).
- Chiu, A. M., Mitra, M., Boymoushakian, L. & Coller, H. A. Integrative analysis of the inter-tumoral heterogeneity of triple-negative breast cancer. *Scientific Reports.* 8, 11807 (2018).

- Garrido-Castro, A. C., Lin, N. & Polyak, K. Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment. *Cancer Discovery.* 9, 176–198 (2019).
- 215. Kao, J., Salari, K., Bocanegra, M., Choi, Y. L., Girard, L., Gandhi, J., Kwei, K. A., Hernandez-Boussard, T., Wang, P., Gazdar, A. F., Minna, J. D. & Pollack, J. R. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLOS One.* 4, e6146 (2009).
- 216. Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J. M., Jacquemier, J., Viens, P., Birnbaum, D. & Bertucci, F. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Research and Treatment. 126, 407–420 (2011).
- 217. Neve, R., Chin, K., Fridlyand, J., Yeh, J., Baehner, F., Fevr, T., Clark, L., Bayani, N., Coppe, J., Tong, F., Speed, T., Spellman, P., DeVries, S., Lapuk, A., Wang, N., Kuo, W., Stilwell, J., Pinkel, D., Albertson, D., Waldman, F., McCormick, F., Dickson, R., Johnson, M., Lippman, M., Ethier, S., Gazdar, A. & Gray, J. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.* 10, 515–527 (2006).
- Polyak, K. Breast cancer: origins and evolution. Journal of Clinical Investigation. 117, 3155–3163 (2007).
- Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology. 15, 81–94 (2018).
- 220. Markowetz, F. & Spang, R. Inferring cellular networks-a review. BMC Bioinformatics.
  27 (2007).
- 221. Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A. & SanCristobal, M. The structure of a gene co-expression network reveals biological functions underlying eQTLs. *PLoS One.* 8, e60045 (2013).
- 222. Zuo, Y., Yu, G., Tadesse, M. G. & Ressom, H. W. Biological network inference using low order partial correlation. *Methods.* 69, 266–73 (2014).

- Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics.* 20, 1746–1758 (2004).
- 224. Bland, K. I., Konstadoulakis, M. M., Vezeridis, M. P. & Wanebo, H. J. Oncogene protein co-expression. Value of Ha-ras, c-myc, c-fos, and p53 as prognostic discriminants for breast carcinoma. *Annals of Surgery.* 221, 706–720 (1995).
- 225. Koedoot, E., Fokkelman, M., Rogkoti, V., Smid, M., van de Sandt, I., de Bont, H., Pont, C., Klip, J. E., Wink, S., Timmermans, M. A., Wiemer, E. A. C., Stoilov, P., Foekens, J. A., Le Devedec, S. E., Martens, J. W. M. & van de Water, B. Uncovering the signaling landscape controlling breast cancer cell migration identifies novel metastasis driver genes. *Nature Communications.* 10, 2983. (2019).
- 226. Tatebe, K., Zeytun, A., Ribeiro, R. M., Hoffmann, R., Harrod, K. S. & Forst, C. V. Response network analysis of differential gene expression in human epithelial lung cells during avian influenza infections. *BMC Bioinformatics.* **11** (2010).
- 227. Tullai, J. W., Schaffer, M. E., Mullenbrock, S., Sholder, G., Kasif, S. & Cooper, G. M. Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *Journal of Biological Chemistry.* 282, 23981–23995 (2007).
- Dunn, K. L., Espino, P. S., Drobic, B., He, S. & Davie, J. R. The Ras-MAPK signal transduction pathway, cancer and chromatin remodeling. *Biochemistry and Cell Biology*.
   83, 1–14 (2005).
- Healy, S., Khan, P. & Davie, J. R. Immediate early response genes and cell transformation.
   Pharmacology and Therapeutics. 137, 64–77 (2013).
- 230. Langer, S., Singer, C. F., Hudelist, G., Dampier, B., Kaserer, K., Vinatzer, U., Pehamberger, H., Zielinski, C., Kubista, E. & Schreibner, M. Jun and Fos family protein expression in human breast cancer: correlation of protein expression and clinicopathological parameters. *European Journal of Gynaecological Oncology.* 27, 345–52 (2006).

- 231. Sacchetti, P., Carpentier, R., Segard, P., Olive-Cren, C. & Lefebvre, P. Multiple signaling pathways regulate the transcriptional activity of the orphan nuclear receptor NURR1. *Nucleic Acids Research* 34, 5515–5527 (2006).
- 232. Wan, P. K., Siu, M. K., Leung, T. H., Mo, X. T., Chan, K. K. & Ngan, H. Y. Role of Nurr1 in Carcinogenesis and Tumor Immunology: A State of the Art Review. *Cancers* 12, 3044 (2020).
- Llopis, S., Singleton, B., Duplessis, T., Carrier, L., Rowan, B. & Williams, C. Dichotomous roles for the orphan nuclear receptor NURR1 in breast cancer. *BMC Cancer.* 13, 139 (2013).
- Jochum, W., Passegue, E. & Wagner, E. F. AP-1 in mouse development and tumorigenesis. Oncogene. 20, 2401–2412 (2001).
- 235. Mikula, M., Gotzmann, J., Fischer, A. N., Wolschek, M. F., Thallinger, C., Schulte-Hermann, R., Beug, H. & Mikulits, W. The proto-oncoprotein c-Fos negatively regulates hepatocellular tumorigenesis. *Oncogene.* 22, 6725–6738 (2003).
- Salomoni, P. & Pandolfi, P. P. The role of PML in tumor suppression. *Cell.* 108, 165–170 (2002).
- 237. Martin-Martin, N., Piva, M., Urosevic, J., Aldaz, P., Sutherland, J. D., Fernandez-Ruiz, S., Arreal, L., Torrano, V., Cortazar, A. R., Planet, E., Guiu, M., Radosevic-Robin, N., Garcia, S., Macias, I., Salvador, F., Domenici, G., Rueda, O. M., Zabala-Letona, A., Arruabarrena-Aristorena, A., Zuniga-Garcia, P., Caro-Maldonado, A., Valcarcel-Jimenez, L., Sanchez-Mosquera, P., Varela-Rey, M., Martinez-Chantar, M. L., Anguita, J., Ibrahim, Y. H., Scaltriti, M., Lawrie, C. H., Aransay, A. M., Iovanna, J. L., Baselga, J., Caldas, C., Barrio, R., Serra, V., Vivanco, M., Matheu, A., Gomis, R. R. & Carracedo, A. Stratification and therapeutic potential of PML in metastatic breast cancer. *Nature Communications* 7 (2016).
- 238. Ponente, M., Campanini, L., Cuttano, R., Piunti, A., Delledonne, G. A., Coltella, N., Valsecchi, R., Villa, A., Cavallaro, U., Pattini, L., Doglioni, C. & Bernardi, R. PML

promotes metastasis of triple-negative breast cancer through transcriptional regulation of HIF1A target genes. *JCI Insight* **2**, e87380 (2017).

- 239. Arreal, L., Piva, M., Fernandez, S., Revandkar, A., Schaub-Clerigue, A., Villanueva, J., Zabala-Letona, A., Pujana, M., Astobiza, I., Cortazar, A. R., Hermanova, I., Bozal-Basterra, L., Arruabarrena-Aristorena, A., Crespo, J. R., Valcarcel-Jimenez, L., Zuniga-Garcia, P., Canals, F., Torrano, V., Barrio, R., Sutherland, J. D., Alimonti, A., Martin-Martin, N. & Carracedo, A. Targeting PML in triple negative breast cancer elicits growth suppression and senescence. *Cell Death and Differentiation* 27, 1186–1199 (2020).
- 240. Hsu, K. S., Zhao, X., Cheng, X., Guan, D., Mahabeleshwar, G. H., Liu, Y., Borden, E., Jain, M. K. & Kao, H. Y. Dual regulation of Stat1 and Stat3 by the tumor suppressor protein PML contributes to interferon-mediated inhibition of angiogenesis. *The Journal* of Biological Chemistry. 292, 10048–10060 (2017).
- 241. Chen, L., Zeng, X., Kleibeuker, E., Buffa, F., Barberis, A., Leek, R. D., Roxanis, I., Zhang, W., Worth, A., Beech, J. S., Harris, A. L. & Cai, S. Paracrine effect of GTP cyclohydrolase and angiopoietin-1 interaction in stromal fibroblasts on tumor Tie2 activation and breast cancer growth. *Oncotarget* 7, 9353–9367 (2016).
- 242. Shimizu, S., Hiroi, T., Ishii, M., Hagiwara, T., Wajima, T., Miyazaki, A. & Kiuchi, Y. Hydrogen peroxide stimulates tetrahydrobiopterin synthesis through activation of the Jak2 tyrosine kinase pathway in vascular endothelial cells. *The International Journal of Biochemistry and Cell Biology* 40, 755–765 (2008).
- 243. Huang, A., Zhang, Y. Y., Chen, K., Hatakeyama, K. & Keaney, J. F. J. Cytokinestimulated GTP cyclohydrolase I expression in endothelial cells requires coordinated activation of nuclear factor-kappaB and Stat1/Stat3. *Circulation Research* 96, 164–171 (2005).
- 244. Mehraj, V., Textoris, J., Ben Amara, A., Ghigo, E., Raoult, D., Capo, C. & Mege, J. L. Monocyte responses in the context of Q fever: from a static polarized model to a kinetic model of activation. *The Journal of Infectious Diseases.* 208, 942–951 (2013).

- 245. Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., Mose, L. E., Vo, N. S., Liu, J., Liu, Y., Rader, J., Dhankani, V., Reynolds, S. M., Bowlby, R., Califano, A., Cherniack, A. D., Anastassiou, D., Bedognetti, D., Mokrab, Y., Newman, A. M., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T. M., Noushmehr, H., Pedamallu, C. S., Bullman, S., Ojesina, A. I., Lamb, A., Zhou, W., Shen, H., Choueiri, T. K., Weinstein, J. N., Guinney, J., Saltz, J., Holt, R. A., Rabkin, C. S., Lazar, A. J., Serody, J. S., Demicco, E. G., Disis, M. L., Vincent, B. G. & Shmulevich, I. The Immune Landscape of Cancer. *Immunity.* 48, 812–830 (2018).
- Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. Nature Reviews Immunology. 14, 36–49 (2014).
- 247. Pascual-Garcia, M., Rue, L., Leon, T., Julve, J., Carbo, J. M., Matalonga, J., Auer, H., Celada, A., Escola-Gil, J. C., Steffensen, K. R., Perez-Navarro, E. & Valledor, A. F. Reciprocal negative cross-talk between liver X receptors (LXRs) and STAT1: effects on IFN-γ-induced inflammatory responses and LXR-dependent gene expression. *Journal of Immunology.* **190**, 6520–6532 (2013).
- Barab asi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics.* 5, 101–113. (2004).
- 249. Furlong, L. I. Human diseases through the lens of network biology. Trends in Genetics.
  29, 150–159. (2013).
- Cho, D. Y., Kim, Y. A. & Przytycka, T. M. Chapter 5: Network biology approach to complex diseases. *PLOS Computational Biology.* 8, e1002820 (2012).
- Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics.* 17, 615–629 (2016).
- 252. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Briefings in Bioinformatics.* 19, 1370–1381 (2018).

- 253. Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J.
  & Perou, C. M. Concordance among gene-expression-based predictors for breast cancer. The New England Journal of Medicine. 355 (2006).
- 254. Bertucci, F., Finetti, P., Cervera, N., Esterni, B., Hermitte, F., Viens, P. & Birnbaum,
  D. How basal are triple-negative breast cancers? *International Journal of Cancer.* 123, 236–240 (2008).
- 255. Yehiely, F., Moyano, J. V., Evans, J. R., Nielsen, T. O. & Cryns, V. L. Deconstructing the molecular portrait of basal-like breast cancer. *Trends in Molecular Medicine*. **12**, 537–544 (2006).
- 256. Schneider, B. P., Winer, E. P., Foulkes, W. D., Garber, J., Perou, C. M., Richardson, A., Sledge, G. W. & Carey, L. A. Triple-negative breast cancer: risk factors to potential targets. *Clinical Cancer Research.* 14, 8010–8018 (2008).
- 257. Wahba, H. A. & El-Hadaad, H. A. Current approaches in treatment of triple-negative breast cancer. *Cancer Biology and Medicine*. **12**, 106–116. (2015).
- 258. Linderholm, B. K., Hellborg, H., Johansson, U., Elmberger, G., Skoog, L., Lehtio, J. & Lewensohn, R. 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 (2009).
- 259. Miller, K., Wang, M., Gralow, J., Dickler, M., Cobleigh, M., Perez, E. A., Shenkier, T., Cella, D. & Davidson, N. E. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *The New England Journal of Medicine*. 357, 2666–2676. (2007).
- 260. Vahdat, L. T., Layman, R., Yardley, D. A., Gradishar, W., Salkeni, M. A., Joy, A. A., Garcia, A. A., Ward, P., Khatcheressian, J., Sparano, J., Rodriguez, G., Tang, S., Gao, L., Dalal, R. P., Kauh, J. & Miller, K. Randomized Phase II Study of Ramucirumab or Icrucumab in Combination with Capecitabine in Patients with Previously Treated Locally Advanced or Metastatic Breast Cancer. Oncologist. 22, 245–254 (2017).

- 261. Zafrakas, M., Papasozomenou, P. & Emmanouilides, C. Sorafenib in breast cancer treatment: A systematic review and overview of clinical trials. World Journal of Clinical Oncology. 7, 331–336. (2016).
- 262. Nakhjavani, M., Hardingham, J. E., Palethorpe, H. M., Price, T. J. & Townsend, A. R. Druggable Molecular Targets for the Treatment of Triple Negative Breast Cancer. *Journal* of Breast Cancer. 22, 341–361. (2019).
- 263. Bohn, O. L., Fuertes-Camilo, M., Navarro, L., Saldivar, J. & Sanchez-Sosa, S. p16INK4a expression in basal-like breast carcinoma. *International Journal of Clinical and Experimental Pathology.* 3, 600–607 (2010).
- 264. Masuda, H., Zhang, D., Bartholomeusz, C., Doihara, H., Hortobagyi, G. N. & Ueno, N. T. Role of epidermal growth factor receptor in breast cancer. *Breast Cancer Research and Treatment.* 136 (2012).
- 265. Maennling, A. E., Tur, M. K., Niebert, M., Klockenbring, T., Zeppernick, F., Gattenlohner, S., Meinhold-Heerlein, I. & Hussain, A. F. Molecular Targeting Therapy against EGFR Family in Breast Cancer: Progress and Future Potentials. *Cancers (Basel).* 11, 1826 (2019).
- 266. Finn, R. S., Press, M. F., Dering, J., Arbushites, M., Koehler, M., Oliva, C., Williams, L. S. & Di Leo, A. 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 (2009).
- 267. Baselga, J., Gomez, P., Greil, R., Braga, S., Climent, M. A., Wardley, A. M., Kaufman, B., Stemmer, S. M., Pego, A., Chan, A., Goeminne, J. C., Graas, M. P., Kennedy, M. J., Ciruelos Gil, E. M., Schneeweiss, A., Zubel, A., Groos, J., Melezinkova, H. & Awada, A. 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 (2013).

- 268. O'Shaughnessy, J., Weckstein, D. J., Vukelja, S. J., McIntyre, K., Krekow, L., Holmes, F. A., Asmar, L. & Blum, J. L. Preliminary results of a randomized phase II study of weekly irinotecan/carboplatin with or without cetuximab in patients with metastatic breast cancer. Breast Cancer Research and Treatment., S32–S33 (2007).
- 269. Massihnia, D., Galvano, A., Fanale, D., Perez, A., Castiglia, M., Incorvaia, L., Listi, A., Rizzo, S., Cicero, G., Bazan, V., Castorina, S. & Russo, A. Triple negative breast cancer: shedding light onto the role of pi3k/akt/mtor pathway. *Oncotarget.* 7, 60712– 60722. (2016).
- 270. Eisen, M., Spellman, P., Brown, P. & Botstein, D. Cluster analysis and display of genomewide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95, 14863–14868 (1998).
- 271. Kim, W., Kim, E., Kim, S., Kim, Y., Ha, Y., Jeong, P., Kim, M., Yun, S., Lee, K., Moon, S., Lee, S., Cha, E. & Bae, S. Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Molecular Cancer* 9, 3 (2010).
- 272. Mouttet, D., Lae, M., Caly, M., Gentien, D., Carpentier, S., Peyro-Saint-Paul, H., Vincent-Salomon, A., Rouzier, R., Sigal-Zafrani, B., Sastre-Garau, X. & Reyal, F. Estrogen-Receptor, Progesterone-Receptor and HER2 Status Determination in Invasive Breast Cancer. Concordance between Immuno-Histochemistry and MapQuant Microarray Based Assay. *PLoS One.* **11**, e0146474 (2016).
- Giulianelli, S., Molinolo, A. & Lanari, C. Targeting progesterone receptors in breast cancer. Vitamins and Hormones. 93, 161–184. (2013).
- Lim, E., Palmieri, C. & Tilley, W. D. Renewed interest in the progesterone receptor in breast cancer. Breast Cancer Research. 115, 909–911. (2016).
- Pernas, S. & Tolaney, S. M. HER2-positive breast cancer: new therapeutic frontiers and overcoming resistance. *Therapeutic Advances in Medical Oncology.* 11 (2019).
- 276. Charitou, T., Bryan, K. & Lynn, D. J. Using biological networks to integrate, visualize and analyze genomics data. *Genetics Selection Evolution*. 48 (2016).

- 277. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D. & Bader, G. D. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols.* 14, 482–517. (2019).
- Pawson, T. & Nash, P. Protein-protein interactions define specificity in signal transduction. Genes and Development. 14, 1027–1047 (2000).
- Ryan, D. P. & Matthews, J. M. Protein-protein interactions in human disease. Current Opinion in Structural Biology. 15, 441–446 (2005).
- Diaz-Eufracio, B. I., Naveja, J. J. & Medina-Franco, J. L. Protein-Protein Interaction Modulators for Epigenetic Therapies. Advances in Protein Chemistry and Structural Biology. 110, 65–84. (2018).
- 281. Goncearenco, A., Li, M., Simonetti, F. L., Shoemaker, B. A. & Panchenko, A. R. Exploring Protein-Protein Interactions as Drug Targets for Anti-cancer Therapy with In Silico Workflows. *Methods in Molecular Biology.* 1647, 221–236 (2017).
- 282. Zhang, G., Andersen, J. & Gerona-Navarro, G. Peptidomimetics Targeting Protein-Protein Interactions for Therapeutic Development. *Protein and Peptide Letters.* 25, 1076–1089 (2018).
- 283. Robertson, N. S. & Spring, D. R. Using Peptidomimetics and Constrained Peptides as Valuable Tools for Inhibiting ProteinProtein Interactions. *Molecules.* 23, 959 (2018).
- 284. Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N. & Liu, E. A. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*. **303**, 844–848. (2004).
- 285. Zhao, Y., Aguilar, A., Bernard, D. & Wang, S. Small-molecule inhibitors of the MDM2p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment.
  J. Journal of Medicinal Chemistry. 58, 1038–1052. (2015).

- Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nature Reviews Drug Discovery*. 15, 533–550. (2016).
- 287. Zinzalla, G. & Thurston, D. E. Targeting protein-protein interactions for therapeutic intervention: a challenge for the future. *Future Medicinal Chemistry.* 1, 65–93 (2009).
- Verdine, G. L. & Walensky, L. D. The challenge of drugging undruggable targets in cancer: lessons learned from targeting BCL-2 family members. *Clinical Cancer Research*. 13, 7264–7270 (2007).
- Murray, J. K. & Gellman, S. H. Targeting protein-protein interactions: lessons from p53/Mdm2. Journal of Peptide Science. 88 (2007).
- 290. Meche, A., Cimpean, A. M. & Raica, M. Immunohistochemical expression and significance of epidermal growth factor receptor (EGFR) in breast cancer. *Romanian Journal* of Morphology and Embryology. 50, 217–221 (2009).
- 291. 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 (2016).
- 292. Holbro, T., Beerli, R. R., Maurer, F., Koziczak, M., Barbas, C. F. & Hynes, N. E. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. *Proceedings of the National Academy of Sciences of the* United States of America. 100, 8933–8938. (2003).
- 293. Lee-Hoeflich, S. T., Crocker, L., Yao, E., Pham, T., Munroe, X., Hoeflich, K. P., Sliwkowski, M. X. & Stern, H. M. A central role for HER3 in HER2-amplified breast cancer: implications for targeted therapy. *Cancer Research.* 68, 5878–5887. (2008).
- Lee, Y., Ma, J., Lyu, H., Huang, J., Kim, A. & Liu, B. Role of erbB3 receptors in cancer therapeutic resistance. Acta Biochimica et Biophysica Sinica(Shanghai). 46, 190–198. (2014).

- 295. Amin, D. N., Campbell, M. R. & Moasser, M. M. The role of HER3, the unpretentious member of the HER family, in cancer biology and cancer therapeutics. *Seminars in Cell* and Developmental Biology. 21, 944–950 (2010).
- 296. Ma, J., Lyu, H., Huang, J. & Liu, B. Targeting of erbB3 receptor to overcome resistance in cancer treatment. *Molecular Cancer.* 13 (2014).
- 297. Lyu, H., Han, A., Polsdofer, E., Liu, S. & Liu, B. Understanding the biology of HER3 receptor as a therapeutic target in human cancer. Acta Pharmaceutica Sinica B. 8, 503– 510 (2018).
- 298. Mishra, R., Patel, H., Alanazi, S., Yuan, L. & Garrett, J. T. HER3 signaling and targeted therapy in cancer. Oncology Reviews. 12, 355 (2018).
- 299. Luhtala, S., Staff, S., Kallioniemi, A., Tanner, M. & Isola, J. Clinicopathological and prognostic correlations of HER3 expression and its degradation regulators, NEDD4-1 and NRDP1, in primary breast cancer. *BMC Cancer.* 18, 1045 (2018).
- 300. Cantley, L. The phosphoinositide 3-kinase pathway. Science. 296, 1655–1657 (2002).
- LoRusso, P. Inhibition of the PI3K/AKT/mTOR Pathway in Solid Tumors. Journal of Clinical Oncology 34, 3803–3815 (2016).
- 302. Chan, J., Tan, T. & Dent, R. Novel therapeutic avenues in triple-negative breast cancer: PI3K/AKT inhibition, androgen receptor blockade, and beyond. *Therapeutic Advances in Medical Oncology* **11** (2019 Oct 9).
- 303. Yarden, Y., Kuang, W., Yang-Feng, T., Coussens, L., Munemitsu, S., Dull, T., Chen, E., Schlessinger, J., Francke, U. & Ullrich, A. Human proto-oncogene c-kit: a new cell surface receptor tyrosine kinase for an unidentified ligand. *The EMBO Journal* 6, 3341– 3351 (1987).
- 304. Tsujimura, T., Hashimoto, K., Kitayama, H., Ikeda, H., Sugahara, H., Matsumura, I., Kaisho, T., Terada, N., Kitamura, Y. & Kanakura, Y. Activating mutation in the catalytic domain of c-kit elicits hematopoietic transformation by receptor self-association not at the ligand-induced dimerization site. *Blood.* 93, 1319–1329 (1999).

- 305. Casteran, N., De Sepulveda, P., Beslu, N., Aoubala, M., Letard, S., Lecocq, E., Rottapel,
  R. & Dubreuil, P. Signal transduction by several KIT juxtamembrane domain mutations.
  Oncogene. 22, 4710–4722 (2003).
- 306. Piao, X., Paulson, R., van der Geer, P., Pawson, T. & A., B. Oncogenic mutation in the Kit receptor tyrosine kinase alters substrate specificity and induces degradation of the protein tyrosine phosphatase SHP-1. Proceedings of the National Academy of Sciences of the United States of America 93, 14665–14669. (1996).
- 307. Naoe, T. & Kiyoi, H. Normal and oncogenic FLT3. Cellular and Molecular Life Sciences
  61, 2932–2938 (2004).
- 308. Miettinen, M. & Lasota, J. KIT (CD117): a review on expression in normal and neoplastic tissues, and mutations and their clinicopathologic correlation. Applied Immunohistochemistry and Molecular Morphology. 13, 205–220 (2005).
- 309. Jr. Roskoski, R. Structure and regulation of Kit protein-tyrosine kinase-the stem cell factor receptor. *Biochemical and Biophysical Research Communications.* 338, 1307–1315 (2005).
- DiPaola, R. S., Kuczynski, W. I., Onodera, K., Ratajczak, M. Z., Hijiya, N., Moore, J. & Gewirtz, A. M. Evidence for a functional kit receptor in melanoma, breast, and lung carcinoma cells. *Cancer Gene Therapy.* 4, 176–182 (1997).
- 311. Simon, R., Panussis, S., Maurer, R., Spichtin, H., Glatz, K., Tapia, C., Mirlacher, M., Rufle, A., Torhorst, J. & Sauter, G. KIT (CD117)-positive breast cancers are infrequent and lack KIT gene mutations. *Clinical Cancer Research.* 10, 178–183 (2004).
- 312. Tsutsui, S., Yasuda, K., Suzuki, K., Takeuchi, H., Nishizaki, T., Higashi, H. & Era, S. A loss of c-kit expression is associated with an advanced stage and poor prognosis in breast cancer. Breast Cancer Research. 94, 1874–1878 (2006).
- 313. Nalwoga, H., Arnes, J., Wabinga, H. & Akslen, L. Expression of EGFR and c-kit is associated with the basal-like phenotype in breast carcinomas of African women. APMIS. 116 (2008).

- 314. Johansson, I., Aaltonen, K., Ebbesson, A., Grabau, D., Wigerup, C., Hedenfalk, I. & Ryden, L. Increased gene copy number of KIT and VEGFR2 at 4q12 in primary breast cancer is related to an aggressive phenotype and impaired prognosis. *Genes Chromosomes Cancer.* 51 (2012).
- 315. Kashiwagi, S., Yashiro, M., Takashima, T., Aomatsu, N., Kawajiri, H., Ogawa, Y., Onoda, N., Ishikawa, T., Wakasa, K. & Hirakawa, K. c-Kit expression as a prognostic molecular marker in patients with basal-like breast cancer. *British Journal of Surgery.* 100(4) (2013).
- 316. Kim, M., Ro, J., Ahn, S., Kim, H., Kim, S. & Gong, G. Clinicopathologic significance of the basal-like subtype of breast cancer: a comparison with hormone receptor and Her2/neu-overexpressing phenotypes. *Human Pathology.* 37 (2006).
- 317. Feng, Z., Riopel, M., Popell, A. & Wang, R. A survival Kit for pancreatic beta cells: stem cell factor and c-Kit receptor tyrosine kinase. *Diabetologia* 58, 654–665 (2015).
- 318. Maroun, C., Moscatello, D., Naujokas, M., Holgado-Madruga, M., Wong, A. & Park, M. A conserved inositol phospholipid binding site within the pleckstrin homology domain of the Gab1 docking protein is required for epithelial morphogenesis. A conserved inositol phospholipid binding site within the pleckstrin homology domain of the Gab1 docking protein is required for epithelial morphogenesis. J Biol Chem. 274, 31719–31726 (1999).
- 319. Lock, L., Maroun, C., Naujokas, M. & M., P. Distinct recruitment and function of Gab1 and Gab2 in Met receptor-mediated epithelial morphogenesis. *Mol Biol Cell.* 13, 2132– 2146 (2002).
- 320. Wohrle, F., Daly, R. & Brummer, T. Function, regulation and pathological roles of the Gab/DOS docking proteins. *Cell Commun Signal* 7, 22 (2009).
- 321. Abella, J., Vaillancourt, R., Frigault, M., Ponzo, M., Zuo, D., Sangwan, V., Larose, L. & Park, M. The Gab1 scaffold regulates RTK-dependent dorsal ruffle formation through the adaptor Nck. *Journal of Cell Science* 123, 1306–1319 (2010).

- 322. Seiden-Long, I., Navab, R., Shih, W., Li, M., Chow, J., Zhu, C. Q., Radulovich, N., Saucier, C. & Tsao, M. S. Gab1 but not Grb2 mediates tumor progression in Met overexpressing colorectal cancer cells. *Carcinogenesis.* 29, 647–655 (2008).
- 323. Sang, H., Li, T., Li, H. & Liu, J. Down-regulation of Gab1 inhibits cell proliferation and migration in hilar cholangiocarcinoma. *PLoS One.* 8, e81347 (2013).
- 324. Hoeben, A., Martin, D., Clement, P. M., Cools, J. & Gutkind, J. S. Role of GRB2associated binder 1 in epidermal growth factor receptor-induced signaling in head and neck squamous cell carcinoma. *International Journal of Cancer.* **132** (2013).
- 325. Bai, R., Weng, C., Dong, H., Li, S., Chen, G. & Xu, Z. MicroRNA-409-3p suppresses colorectal cancer invasion and metastasis partly by targeting GAB1 expression. *International Journal of Cancer.* 137 (2015).
- 326. Sun, W., Zhang, Z., Wang, J., Shang, R., Zhou, L., Wang, X., Duan, J., Ruan, B., Gao, Y., Dai, B., Qu, S., Liu, W., Ding, R., Wang, L., Wang, D. & Dou, K. MicroRNA-150 suppresses cell proliferation and metastasis in hepatocellular carcinoma by inhibiting the GAB1-ERK axis. Oncotarget. 7, 11595–11608. (2016).
- 327. Hu, L. & Liu, R. Expression of Gab1 is associated with poor prognosis of patients with epithelial ovarian Cancer. The Tohoku Journal of Experimental Medicine. 239, 177–184. (2016).
- Wang, X., Peng, J., Yang, Z., Zhou, P. J., An, N., Wei, L., Zhu, H. H., Lu, J., Fang, Y. X. & Gao, W. Q. Elevated expression of Gab1 promotes breast cancer metastasis by dissociating the PAR complex. *Journal of Experimental and Clinical Cancer Research.* 38, 27 (2019).
- Granito, A., Guidetti, E. & Gramantieri, L. c-MET receptor tyrosine kinase as a molecular target in advanced hepatocellular carcinoma. *Journal of Hepatocellular Carcinoma* 2, 29– 38 (2015).
- 330. Domchek, S. M., Auger, K. R., Chatterjee, S., Jr Burke, T. R. & Shoelson, S. E. Inhibition of SH2 domain/phosphoprotein association by a nonhydrolyzable phosphonopeptide. *Biochemistry.* **31**, 9865–9870. (1992).

- 331. Herbst, R., Shearman, M. S., Jallal, B., Schlessinger, J. & Ullrich, A. Formation of signal transfer complexes between stem cell and platelet-derived growth factor receptors and SH2 domain proteins in vitro. *Biochemistry.* 34, 5971–5979 (1995).
- 332. Chen, T. C., Lin, K. T., Chen, C. H., Lee, S. A., Lee, P. Y., Liu, Y. W., Kuo, Y. L., Wang, F. S., Lai, J. M. & Huang, C. Y. Using an in situ proximity ligation assay to systematically profile endogenous protein-protein interactions in a pathway network. *Journal of Proteome Research.* 13, 5339–5346. (2014).
- 333. Lehr, S., Kotzka, J., Herkner, A., Sikmann, A., Meyer, H. E., Krone, W. & Muller-Wieland, D. Identification of major tyrosine phosphorylation sites in the human insulin receptor substrate Gab-1 by insulin receptor kinase in vitro. *Biochemistry.* 39, 10898– 10907 (2000).
- 334. Montagner, A., Yart, A., Dance, M., Perret, B., Salles, J. P. & Raynal, P. A novel role for Gab1 and SHP2 in epidermal growth factor-induced Ras activation. *Journal of Biological Chemistry.* 280, 5350–5360. (2005).
- 335. Nakaoka, Y., Nishida, K., Fujio, Y., Izumi, M., Terai, K., Oshima, Y., Sugiyama, S., Matsuda, S., Koyasu, S., Yamauchi-Takihara, K., Hirano, T., Kawase, I. & Hirota, H. Activation of gp130 transduces hypertrophic signal through interaction of scaffolding/docking protein Gab1 with tyrosine phosphatase SHP2 in cardiomyocytes. *Circulation Research.* 93, 221–229. (2003).
- Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Research. 46, D493–D496 (2018).
- Pundir, S., Martin, M. J. & O'Donovan, C. UniProt Protein Knowledgebase. Methods in Molecular Biology. 1558 (2017).
- 338. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. & Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. 7 (2011).

- 339. Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E. & Gibson, T. J. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Molecular BioSystems.* 10, 2626–2642 (2014).
- 340. Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., Gibson, T. J. & Davey, N. E. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chemical Reviews.* 114, 6733–6778. (2014).
- 341. Gouw, M., Michael, S., Samano-Sanchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., Gurth, C. M., Huber, A. K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg, B., Remenyi, A., Dinkel, H. & Gibson, T. J. The eukaryotic linear motif resource 2018 update. *Nucleic Acids Research*. 46, D428–D434. (2018).
- 342. Schad, E., Ficho, E., Pancsa, R., Simon, I., Dosztanyi, Z., Meszaros, B., Schad, E., Ficho, E., Pancsa, R., Simon, I., Dosztanyi, Z. & Meszaros, B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics.* 34, 535–537. (2018).
- 343. Edfors, F., Danielsson, F., Hallstrom, B. M., Kall, L., Lundberg, E., Ponten, F., Forsstrom,
  B. & Uhlen, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology.* 12, 883 (2016).
- 344. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature.* 547, E19–E20 (2017).
- 345. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics.* **13**, 227–32 (2012).
- 346. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature*. 473, 337–342 (2011).
- 347. Sharova, L. V., Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N. & Ko, M. S. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Research. 16, 45–58 (2009).

- 348. Febbo, P. G. & Kantoff, P. W. Noise and bias in microarray analysis of tumor specimens. Journal of Clinical Oncology. 24, 3719–21 (2006).
- 349. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology.* 4, 117 (2003).
- 350. Kustatscher, G., Grabowski, P. & Rappsilber, J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Molecular Systems Biology.* **13**, 937 (2017).
- 351. Garvin, A., Pawar, S., Marth, J. & Perlmutter, R. Structure of the murine lck gene and its rearrangement in a murine lymphoma cell line. *Molecular and Cellular Biology* 8, 3058–3064 (1988).
- 352. Paillard, F. & Vaquero, C. Down-regulation of lck mRNA by T cell activation involves transcriptional and post-transcriptional mechanisms. *Nucleic Acids Research* 19, 4655– 4661 (1991).
- 353. Marth, J., Lewis, D., Wilson, C., Gearn, M., Krebs, E. & Perlmutter, R. Regulation of p56lck during T-cell activation: functional implications for the src-like protein tyrosine kinases. *The EMBO Journal* 6, 2727–2734 (1987).
- 354. He, F., Ge, W., Martinowich, K., Becker-Catania, S., Coskun, V., Zhu, W., Wu, H., Castro, D., Guillemot, F., Fan, G., de Vellis, J. & YE., S. A positive autoregulatory loop of Jak-STAT signaling controls the onset of astrogliogenesis. *Nature Neuroscience* 8, 616–625 (2005).
- 355. Yuasa, K. & Hijikata, T. Distal regulatory element of the STAT1 gene potentially mediates positive feedback control of STAT1 expression. *Genes Cells.* **21**, 25–40 (2016).
- Schindler, C., Levy, D. & T., D. JAK-STAT signaling: from interferons to cytokines. Journal of Biological Chemistry 282, 20059–20063 (2007s).
- 357. Wesoly, J. & Szweykowska-Kulinska, Z. a. B. H. STAT activation and differential complex formation dictate selectivity of interferon responses. Acta Biochimica Polonica 54, 27–38 (2007).

- 358. Levy, D., Kessler, D., Pine, R. & Darnell, J. J. Cytoplasmic activation of ISGF3, the positive regulator of interferon-alpha-stimulated transcription, reconstituted in vitro. Genes and Development 3, 1362–1371 (1989).
- 359. Fu, X., Kessler, D., Veals, S., Levy, D. & Jr., D. J. ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains. *Proceed*ings of the National Academy of Sciences of the United States of America 87, 8555–8559 (1990).
- 360. Levy, D., Kessler, D., Pine, R., Reich, N. & Darnell, J. J. Interferon-induced nuclear factors that bind a shared promoter element correlate with positive and negative transcriptional control. *Genes and Development* 2, 383–393 (1988).
- 361. Darnell JE Jr Kerr IM, S. G. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264**, 1415–1421 (1994).
- 362. Tang, W., Zhou, M., Dorsey, T., Prieto, D., Wang, X., Ruppin, E., Veenstra, T. & Ambs, S. Integrated proteotranscriptomics of breast cancer reveals globally increased proteinmRNA concordance associated with subtypes and survival. *Genome Medicine* 10, 94 (2018).
- 363. Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. & Ideker, T. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics.* 7 (2006).
- Bader, G., Cary, M. & C., S. Pathguide: a pathway resource list. Nucleic Acids Res 34 (2006).
- 365. Yates, J. Recent technical advances in proteomics. 1000Res. 8 (2019).
- 366. Gyorffy, B., Lanczky, A., Eklund, A., Denkert, C., Budczies, J., Li, Q. & Szallasi, Z. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment* 123, 725–731 (2010).
- 367. Nagy A Munkarcsy G, G. B. Pancancer survival analysis of cancer hallmark genes. Scientific Reports 11, 6047 (2021).

- 368. Pawson, T. Protein modules and signalling networks. Nature. 373, 573–580. (1995).
- 369. O'Brien, R., Rugman, P., Renzoni, D., Layton, M., Handa, R., Hilyard, K., Waterfield, M. D., Driscoll, P. C. & Ladbury, J. E. Alternative modes of binding of proteins with tandem SH2 domains. *Protein Science*. 9, 570–579 (2000).
- 370. Ottinger, E. A., Botfield, M. C. & Shoelson, S. E. Tandem SH2 domains confer high specificity in tyrosine kinase signaling. *Journal of Biological Chemistry.* 273, 729–35 (1998).
- 371. Yan, Q., Barros, T., Visperas, P. R., Deindl, S., Kadlecek, T. A., Weiss, A. & Kuriyan, J. Structural basis for activation of ZAP-70 by phosphorylation of the SH2-kinase linker. Molecular and Cellular Biology. 33, 2188–2201. (2013).
- 372. Hayashi, T., Senda, M., Suzuki, N., Nishikawa, H., Ben, C., Tang, C., Nagase, L., Inoue, K., Senda, T. & Hatakeyama, M. Differential Mechanisms for SHP2 Binding and Activation Are Exploited by Geographically Distinct Helicobacter pylori CagA Oncoproteins. *Cell Reports.* 20, 2876–2890 (2017).
- 373. Yu, J., Wjasow, C. & Backer, J. M. Regulation of the p85/p110alpha phosphatidylinositol 3'-kinase. Distinct roles for the n-terminal and c-terminal SH2 domains. *Journal of Biological Chemistry.* 273, 30199–30203 (1998).
- 374. Feng, C., Roy, A. & Post, C. B. Entropic allostery dominates the phosphorylationdependent regulation of Syk tyrosine kinase release from immunoreceptor tyrosine-based activation motifs. *Protein Science.* 27, 1780–1796 (2018).
- Nooren, I. M. & Thornton, J. M. Diversity of protein-protein interactions. *The EMBO Journal.* 22, 3486–3492 (2003).
- Zhang, A. Protein Interaction Networks: Computational Analysis (Cambridge University Press, 2009).
- 377. Monod, J., Wyman, J. & Changeux, J. On the nature of allosteric transitions: a plausible model. Journal of Molecular Biology. 12, 88–118 (1965).
- 378. Hashimoto, K. & Panchenko, A. R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proceedings of*

the National Academy of Sciences of the United States of America. **107**, 20352–20357. (2010).

- 379. Sowmya, G., Breen, E. J. & Ranganathan, S. Linking structural features of protein complexes and biological function. *Protein Science.* **24**, 1486–1494. (2015).
- 380. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical Reviews.* 108, 1225–1244. (2008).
- 381. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 93, 13–20 (1996).
- 382. Sheinerman, F. B., Norel, R. & Honig, B. Electrostatic aspects of protein-protein interactions. Current Opinion in Structural Biology. 10, 153–159 (2000).
- 383. Block, P., Paern, J., Hullermeier, E., Sanschagrin, P., Sotriffer, C. A. & Klebe, G. Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins.* 65, 607–622. (2006).
- 384. Levy, E. D. & Pereira-Leal, J. B. Evolution and dynamics of protein interactions and networks. *Current Opinion in Structural Biology.* 18, 349–357 (2008).
- 385. Nooren, I. M. & Thornton, J. M. Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology.* 325, 991–1018 (2003).
- Janin, J., Bahadur, R. P. & Chakrabarti, P. Protein-protein interaction and quaternary structure. *Quarterly Reviews of Biophysics.* 41, 133–180. (2008).
- 387. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry*. 41, 6573–6582. (2002).
- Dyson, H. & Wright, P. Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology. 6, 197–208 (2005).

- 389. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Letters. 579, 3346–3354 (2005).
- 390. Minezaki, Y., Homma, K., Kinjo, A. R. & Nishikawa, K. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *Journal of Molecular Biology.* 359, 1137–1149 (2006).
- 391. Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T. & Nishikawa, K. Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Structural Biology.* 11 (2011).
- 392. Pancsa, R. & Tompa, P. Structural disorder in eukaryotes. PLoS One. 7, e34687 (2012).
- 393. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal* of Biomolecular Structure and Dynamics. **30**, 137–149. (2012).
- Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*.
   323, 573–584 (2002).
- 395. Anbo, H., Sato, M., Okoshi, A. & Fukuchi, S. Functional Segments on Intrinsically Disordered Regions in Disease-Related Proteins. *Biomolecules.* 9, 88 (2019).
- 396. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology.* 21, 432–440 (2011).
- 397. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews Molecular Cell Biology. 16, 18–29 (2015).
- 398. Rogers, J. M., Oleinikovas, V., Shammas, S. L., Wong, C. T., De Sancho, D., Baker, C. M. & Clarke, J. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proceedings of the National Academy of Sciences of the United States of America.* 111, 15420–15425 (2014).

- 399. Qian, J., Voorbach, M. J., Huth, J. R., Coen, M. L., Zhang, H., Ng, S. C., Comess, K. M., Petros, A. M., Rosenberg, S. H., Warrior, U. & Burns, D. J. Discovery of novel inhibitors of Bcl-xL using multiple high-throughput screening platforms. *Analytical Biochemistry*. 328, 131–8 (2004).
- 400. Real, P. J., Cao, Y., Wang, R., Nikolovska-Coleska, Z., Sanz-Ortiz, J., Wang, S. & Fernandez-Luna, J. L. Breast cancer cells can evade apoptosis-mediated selective killing by a novel small molecule inhibitor of Bcl-2. *Cancer Research.* 64, 7947–7953. (2004).
- 401. Braisted, A. C., Oslob, J. D., Delano, W. L., Hyde, J., McDowell, R. S., Waal, N., Yu, C., Arkin, M. R. & Raimundo, B. C. Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. *Journal of the American Chemical Society.* **125**, 3714–3715 (2003).
- 402. Emerson, S. D., Palermo, R., Liu, C. M., Tilley, J. W., Chen, L., Danho, W., Madison, V. S., Greeley, D. N., Ju, G. & Fry, D. C. NMR characterization of interleukin-2 in complexes with the IL-2Ralpha receptor component, and with low molecular weight compounds that inhibit the IL-2/IL-Ralpha interaction. *Protein Science*. **12**, 811–822 (2003).
- 403. Sadowski, I., Stone, J. C. & Pawson, T. A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. *Molecular and Cellular Biology.* 6, 4396–4408 (1986).
- 404. Stone, J. C., Atkinson, T., Smith, M. & Pawson, T. Identification of functional regions in the transforming protein of Fujinami sarcoma virus by in-phase insertion mutagenesis. *Cell.* 37, 549–558 (1984).
- 405. Mayer, B. J., Hamaguchi, M. & Hanafusa, H. A novel viral oncogene with structural similarity to phospholipase C. Nature. 332, 272–275 (1988).
- 406. Matsuda, M., Mayer, B. J., Fukui, Y. & Hanafusa, H. Binding of transforming protein, P47gag-crk, to a broad range of phosphotyrosine-containing proteins. *Science*. 248, 1537– 1539 (1990).

- 407. Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science.* **300**, 445–452. (2003).
- 408. Morlacchi, P., Robertson, F. M., Klostergaard, J. & McMurray, J. S. Targeting SH2 domains in breast cancer. *Future Medicinal Chemistry.* 6, 1909–1926 (2014).
- 409. Liu, B. A., Jablonowski, K., Raina, M., Arce, M., Pawson, T. & Nash, P. D. The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling. *Molecular Cell.* 22, 851–868. (2006).
- Lappalainen, I., Thusberg, J., Shen, B. & Vihinen, M. Genome wide analysis of pathogenic SH2 domain mutations. *Proteins.* 72, 779–792 (2008).
- Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., Resh, M. D., Rios, C. B., Silverman, L. & Kuriyan, J. Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature.* 358, 646–653 (1992).
- 412. Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell.* 72, 779–790. (1993).
- 413. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins.* **78**, 2029–2040 (2010).
- 414. Jeffrey, G. A. An Introduction to Hydrogen Bonding (Oxford University Press, 1977).
- 415. Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G., King, F., Roberts, T., Ratnofsky, S., Lechleider, R. J., Neel, B. G., Birge, R. B., J. Fajardo, E., Chou, M. M., Hanafusa, H., Schaffhausen, B. & Cantley, L. C. SH2 domains recognize specific phosphopeptide sequences. *Cell.* **72**, 767–778 (1993).
- 416. Visconti, L., Toto, A., Jarvis, J. A., Troilo, F., Malagrino, F., De Simone, A. & Gianni, S. Demonstration of Binding Induced Structural Plasticity in a SH2 Domain. Frontiers in Molecular Biosciences. 7, 89 (2020).

- 417. Kaneko, T., Joshi, R., Feller, S. M. & Li, S. S. Phosphotyrosine recognition domains: the typical, the atypical and the versatile. *Cell Communication and Signaling.* **10**, 32 (2012).
- 418. Liu, B. A., Engelmann, B. W. & Nash, P. D. The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. *FEBS Letters.* 586, 2597– 605 (2012).
- 419. Laffargue, M., Raynal, P., Yart, A., Peres, C., Wetzker, R., Roche, S., Payrastre, B. & Chap, H. An epidermal growth factor receptor/Gab1 signaling pathway is required for activation of phosphoinositide 3-kinase by lysophosphatidic acid. *Journal of Biological Chemistry* 274, 32835–32841 (1999).
- 420. Yart, A., Laffargue, M., Mayeux, P., Chretien, S., Peres, C., Tonks, N., Roche, S., Payrastre, B., Chap, H. & Raynal, P. A critical role for phosphoinositide 3-kinase upstream of Gab1 and SHP2 in the activation of ras and mitogen-activated protein kinases by epidermal growth factor. *Journal of Biological Chemistry* **276**, 8856–8864 (2001).
- 421. Liu, Y. & Rohrschneider, L. The gift of Gab. FEBS Lett. 2002 1-3 (515).
- 422. Gu, H., Maeda, H., Moon, J., Lord, J., Yoakim, M., Nelson, B. & BG., N. New role for Shc in activation of the phosphatidylinositol 3-kinase/Akt pathway. *Molecular and Cellular Biology* 20, 7109–7120 (2000).
- 423. Chin YR, T. A. The actin-bundling protein palladin is an Akt1-specific substrate that regulates breast cancer cell migration. *Molecular Cell* **38**, 333–344 (2010).
- 424. Xu, F., Na, L., Li, Y. & L., C. Roles of the PI3K/AKT/mTOR signalling pathways in neurodegenerative diseases and tumours. *Cell and Bioscience* **10** (2020).
- 425. Sang, H., Li, T., Li, H. & Liu, J. Gab1 regulates proliferation and migration through the PI3K/Akt signaling pathway in intrahepatic cholangiocarcinoma. *Tumor Biology* 36, 8367–8377 (2015).
- 426. De Falco, V., Guarino, V., Malorni, L., Cirafici, A., Troglio, F., Erreni, M., Pelicci, G., Santoro, M. & Melillo, R. RAI(ShcC/N-Shc)-dependent recruitment of GAB 1 to RET

oncoproteins potentiates PI 3-K signalling in thyroid tumors. Oncogene. 24, 6303–6313 (2005).

- 427. Blume-Jensen, P., Janknecht, R. & Hunter, T. The kit receptor promotes cell survival via activation of PI 3-kinase and subsequent Akt-mediated phosphorylation of Bad on Ser136. *Current Biology* 8, 779–782 (1998).
- 428. Serve, H., Hsu, Y. & Besmer, P. Tyrosine residue 719 of the c-kit receptor is essential for binding of the P85 subunit of phosphatidylinositol (PI) 3-kinase and for c-kit-associated PI 3-kinase activity in cos-1 cells. *Journal of Biological Chemistry* 269 (1994).
- 429. Serve, H., Yee, N. S., Stella, G., Sepp-Lorenzino, L., Tan, J. & Besmer, P. Differential roles of PI3-kinase and Kit tyrosine 821 in Kit receptor-mediated proliferation, survival and cell adhesion in mast cells. *The EMBO Journal* 14, 473–483 (1995).
- 430. Kissel, H., Timokhina, I., Hardy, M., Rothschild, G., Tajima, Y., Soares, V., Angeles, M., Whitlow, S., Manova, K. & Besmer, P. Point mutation in kit receptor tyrosine kinase reveals essential roles for kit signaling in spermatogenesis and oogenesis without affecting other kit responses. *The EMBO Journal* 19, 1312–1326 (2000).
- 431. Chian, R., Young, S., Danilkovitch-Miagkova, A., Ronnstrand, L., Leonard, E., Ferrao, P., Ashman, L. & Linnekin, D. Phosphatidylinositol 3 kinase contributes to the transformation of hematopoietic cells by the D816V c-Kit mutant. *Blood* 98, 1365–1373 (2001).
- 432. Shivakrupa, R., Bernstein, A., Watring, N. & Linnekin, D. Phosphatidylinositol 3'-kinase is required for growth of mast cells expressing the kit catalytic domain mutant. *Cancer Research* 63 (2003).
- 433. Hashimoto, K., Matsumura, I., Tsujimura, T., Kim, D., Ogihara, H., Ikeda, H., Ueda, S., Mizuki, M., Sugahara, H., Shibayama, H., Kitamura, Y. & Kanakura, Y. Necessity of tyrosine 719 and phosphatidylinositol 3'-kinase-mediated signal pathway in constitutive activation and oncogenic potential of c-kit receptor tyrosine kinase with the Asp814Val mutation. *Blood.* **101**, 1094–1102 (2003).
- Lennartsson, J. & Ronnstrand, L. Stem cell factor receptor/c-Kit: from basic science to clinical implications. *Physiol Rev.* 92, 1619–1649 (2012).

- 435. Ahmed, M. H., Spyrakis, F., Cozzini, P., Tripathi, P. K., Mozzarelli, A., Scarsdale, J. N., Safo, M. A. & Kellogg, G. E. Bound water at protein-protein interfaces: partners, roles and hydrophobic bubbles as a conserved motif. *PLoS One.* 6, e24712 (2011).
- 436. Parikh, H. I. & Kellogg, G. E. Intuitive, but not simple: including explicit water molecules in protein-protein docking simulations improves model quality. *Proteins.* 82, 916–32 (2014).
- 437. Leroux Vand Gresh, N., Liu, W. Q., Garbay, C. & Maigret, B. Role of water molecules for binding inhibitors in the SH2 domain of Grb2: A molecular dynamics study. *Journal* of Molecular Structure: THEOCHEM. 806, 51–66 (2007).
- 438. Geroult, S., Hooda, M., Virdee, S. & Waksman, G. Prediction of solvation sites at the interface of Src SH2 domain complexes using molecular dynamics simulations. *Chemical Biology and Dug Design.* 70, 87–99 (2007).
- 439. Papoian GA Ulander J, W. P. Role of water mediated interactions in protein-protein recognition landscapes. *Journal of the American Chemical Society*. **125**, 9170–9178 (2003).
- 440. Songyang, Z., Shoelson, S. E., McGlade, J., Olivier, P., Pawson, T., Bustelo, X. R., Barbacid, M., Sabe, H., Hanafusa, H. & Yi, T. Specific motifs recognized by the SH2 domains of Csk, 3BP2, fps/fes, GRB-2, HCP, SHC, Syk, and Vav. *Molecular and Cellular Biology.* 14, 2777–2785 (1994).
- Liu, B. A., Jablonowski, K., Shah, E. E., Engelmann, B. W., Jones, R. B. & Nash, P. D.
   SH2 domains recognize contextual peptide sequence information to determine selectivity. Mol Cell Proteomics. 9, 2391–2404 (2010).
- 442. Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O. & Kmiecik, S. Protein-peptide docking: opportunities and challenges. *Drug Discovery Today.* 23, 1530–1537 (2018).
- Cheng, H., Qi, R., Paudel, H. & Zhu, H. Regulation and function of protein kinases and phosphatases. *Enzyme Research.* **794089** (2011).

- 444. Shchemelinin, I., Sefc, L. & Necas, E. Protein kinases, their function and implication in cancer and other diseases. *Folia Biologica*. **52** (2006).
- 445. Ubersax, J. & Ferrell Jr., J. Mechanisms of specificity in protein phosphorylation. Nature Reviews Molecular Cell Biology. 8 (2007).
- 446. Blume-Jensen, P. & Hunter, T. Oncogenic kinase signalling. Nature. 411 (2001).
- 447. Tsintarakis, A. & Zafiropoulos, A. in, 1–7 (Sept. 2017). ISBN: 9780470016176.
- 448. Cicenas, J., Zalyte, E., Bairoch, A. & Gaudet, P. Kinases and Cancer. Cancers. 10(3) (2018).
- 449. Bhullar, K. S., Lagaron, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P. & Rupasinghe, H. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer.* 17(1) (2018).
- 450. Knapp, S. New opportunities for kinase drug repurposing and target discovery. British Journal of Cancer volume. 118 (2018).
- Midland, A., Whittle, M., Duncan, J., Abell, A., Nakamura, K., Zawistowski, J., Carey, L., Earp, H., Graves, L., Gomez, S. & Johnson, G. Defining the expressed breast cancer kinome. *Cell Research.* 22 (2012).
- 452. Segovia-Mendoza, M., Gonzalez-Gonzalez, M., Barrera, D., Diaz, L. & Garcia-Becerra, R. Efficacy and mechanism of action of the tyrosine kinase inhibitors gefitinib, lapatinib and neratinib in the treatment of HER2-positive breast cancer: Preclinical and clinical evidence. American Journal of Cancer Research. 5 (2015).
- 453. Arteaga, C., Sliwkowski, M., Osborne, C., Perez, E., Puglisi, L. & Gianni, L. Treatment of HER2-positive breast cancer: current status and future perspectives. *Nature Reviews Clinical Oncology.* 9 (2012).
- 454. Marty, B., Maire V.and Gravier, E., Rigaill, G., Vincent-Salomon, A., Kappler, M., Lebigot, I., Djelti, F., Tourdes, A., Gestraud, P., Hupe, P., Barillot, E., Cruzalegui, F., Tucker, G., Stern, M., Thiery, J., Hickman, J. & Dubois, T. Frequent PTEN genomic al-

terations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells. *Breast Cancer Research.* **10(6)** (2008).

- 455. Mirzoeva, O., Das, D., Heiser, L. M., Bhattacharya, S., Siwak, D., Gendelman, R., Bayani, N., Wang, N. J., Neve, R. M., Guan, Y., Hu, Z., Knight, Z., Feiler, H. S., Gascard, P., Parvin, B., Spellman, P. T., Shokat, K. M., Wyrobek, A. J., Bissell, M. J., McCormick, F., Kuo, W., Mills, G., Gray, J. & Korn, W. M. Basal Subtype and MAPK/ERK Kinase (MEK)-Phosphoinositide 3-Kinase Feedback Signaling Determine Susceptibility of Breast Cancer Cells to MEK Inhibition. *Cancer Research.* 69(2) (2009).
- 456. Staff, S., Isola, J., Jumppanen, M. & Tanner, M. Aurora-A gene is frequently amplified in basal-like breast cancer. Oncology Reports. 23(2) (2010).
- 457. Zhang, J., Li, n. X., Wu, L., Huang, J., Jiang, W., Kipps, T. & Zhang, S. Aurora B induces epithelial-mesenchymal transition by stabilizing Snail1 to promote basal-like breast cancer metastasis. *Oncogene.* **39** (2020).
- 458. Butti, R., Das, S., Gunasekaran, V., Yadav, A., Kumar, D. & Kundu, G. Receptor tyrosine kinases (RTKs) in breast cancer: signaling, therapeutic implications and challenges. *Molecular Cancer.* 17(1) (2018).
- 459. Roswall, P., Bocci, M., Bartoschek, M., Li, H., Kristiansen, G., Jansson, S., Lehn, S., Sjolund, J., Reid, S., Larsson, C., Eriksson, P., Anderberg, C., Cortez, E., Saal, L., Orsmark-Pietras, C., Cordero, E., Haller, B., Hakkinen, J., Burvenich, I., Lim, E., Orimo, A., Hoglund, M., Ryden, L., Moch, H., Scott, A., Eriksson, U. & Pietras, K. Microenvironmental control of breast cancer subtype elicited through paracrine platelet-derived growth factor-CC signaling. *Nature Medicine.* 24, 463–473 (2018).
- 460. Jansson, S., Aaltonen, K., Bendahl, P., Falck, A., Karlsson, M., Pietras, K. & Ryden, L. The PDGF pathway in breast cancer is linked to tumour aggressiveness, triple-negative subtype and early recurrence. *Breast Cancer Research and Treatment.* **169(2)** (2018).
- 461. Lerma, E., Peiro, G., Ramon, T., Fernandez, S., Martinez, D., Pons, C., Munoz, F., Sabate, J., Alonso, C., Ojeda, B., Prat, J. & Barnadas, A. Immunohistochemical hetero-

geneity of breast carcinomas negative for estrogen receptors, progesterone receptors and Her2/neu (basal-like breast carcinomas). *Modern Pathology.* **20** (2007).

- Manning, B. & Cantley, L. AKT/PKB signaling: navigating downstream. *Cell.* 129 (2007).
- 463. Cortes, I., Sanchez-Ruiz, J., Zuluaga, S., Calvanese, V., Marques, M., Hernandez, C., Rivera, T., Kremer, L., Gonzalez-Garcia, A. & Carrera, A. C. p85b phosphoinositide 3-kinase subunit regulates tumor progression. *Proceedings of the National Academy of Sciences of the United States of America.* 109, 11318–11323 (2012).
- 464. Cariaga-Martinez, A., Cortes, E., Garcia, I., Perez-Garcia, V., Pajares, M., Idoate, M., Redondo-Munoz, J., Anton, I. & Carrera, A. Phosphoinositide 3-kinase p85beta regulates invadopodium formation. *Biology Open.* 3 (2014).
- 465. Thorpe, L., Spangle, J., Ohlson, C., Cheng, H., Roberts, T., Cantley, L. & Zhao, J. PI3Kp110α mediates the oncogenic activity induced by loss of the novel tumor suppressor PI3K-p85α. Proceedings of the National Academy of Sciences of the United States of America. 114, 7095–7100 (2017).
- 466. Holliday, D. & Speirs, V. Choosing the right cell line for breast cancer research. Breast Cancer Research. 13 (2011).
- 467. Leung, Y., Mak, P., Hassan, S. & Ho, S. Estrogen receptor (ER)-β isoforms: A key to understanding ER-β signaling. Proceedings of the National Academy of Sciences of the United States of America. 103, 13162–13167 (2006).
- 468. Sabatier, R., Finetti, P., Guille, A., Adelaide, J., Chaffanet, M., Viens, P., Birnbaum, D. & Bertucci, F. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Molecular Cancer.* 13 (2014).
- 469. Liu, H., Zang, C., Fenner, M., Possinger, K. & Elstner, E. PPARgamma ligands and ATRA inhibit the invasion of human breast cancer cells in vitro. *Breast Cancer Research* and Treatment. **79**, 63–74 (2003).

- 470. Chavez, K., Garimella, S. & Lipkowitz, S. Triple Negative Breast Cancer Cell Lines: One Tool in the Search for Better Treatment of Triple Negative Breast Cancer. Breast Disease.
  32, 35–48 (2010).
- 471. Hines, S. J., Organ, C., Kornstein, M. J. & Krystal, G. W. Coexpression of the c-kit and stem cell factor genes in breast carcinomas. *Cell growth and differentiation* 6, 769–779 (1995).
- 472. Akkiprik, M., Nicorici, D., Cogdell, D., Jia, Y., Hategan, A., Tabus, I., Yli-Harja, O. D., Sahin, A. & Zhang, W. Dissection of signaling pathways in fourteen breast cancer cell lines using reverse-phase protein lysate microarray. *Technology in Cancer Research and Treatment* 5, 543–551 (2006).
- 473. Ou, O., Huppi, K., Chakka, S., Gehlhaus, K., Dubois, W., Patel, J., Chen, J., Mackiewicz, M., Jones, T., Pitt, J., Martin, S., Goldsmith, P., Simmons, J., Mock, B. & Caplen, N. Loss-of-function RNAi screens in breast cancer cells identify AURKB, PLK1, PIK3R1, MAPK12, PRKD2, and PTK6 as sensitizing targets of rapamycin activity. *Cancer Letters* 354, 336–347 (2014).
- 474. Soderberg, O., Leuchowius, K., Gullberg, M., Jarvius, M., Weibrecht, I., Larsson, L. & Landegren, U. Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. *Methods.* 45, 227–32. (2008).
- 475. Zhang, S. & Fedoroff, S. Cellular localization of stem cell factor and c-kit receptor in the mouse nervous system. *Journal of Neuroscience Research* 47, 1–15 (1997s).
- 476. Unni, S., Modi, D., Pathak, S., Dhabalia, J. & Bhartiya, D. Stage-specific localization and expression of c-kit in the adult human testis. *Journal of Histochemistry and Cytochemistry* 57, 861–869 (2009).
- 477. Tuck, A., Robker, R., Norman, R., Tilley, W. & Hickey, T. Expression and localisation of c-kit and KITL in the adult human ovary. *Journal of Ovarian Research* 8, 31 (2015).
- 478. Lammie, A., Drobnjak, M., Gerald, W., Saad, A., Cote, R. & Cordon-Cardo, C. Expression of c-kit and kit ligand proteins in normal human tissues. *Journal of Histochemistry and Cytochemistry* 42, 1417–1425 (1994).

- 479. Paronetto, M., Farini, D., Sammarco, I., Maturo, G., Vespasiani, G., Geremia, R., Rossi,
  P. & Sette, C. Expression of a truncated form of the c-Kit tyrosine kinase receptor and activation of Src kinase in human prostatic cancer. *The American Journal of Pathology* 164, 1243–1251 (2004).
- 480. Broudy, V., Kovach, N., Bennett, L., Lin, N., Jacobsen, F. & Kidd, P. Human umbilical vein endothelial cells display high-affinity c-kit receptors and produce a soluble form of the c-kit receptor. *Blood.* 83, 2145–2152 (1994).
- 481. Yang, J., Saltiel, C., Nachtman, R., Jing, X. & Jurecic, R. The Role of Truncated c-kit Receptor (tr-kit) in Maintenance and Differentiation of Hematopoietic Stem Cells and Multipotent Hematopoietic Progenitors. *Blood* **106**, 4193 (2005).
- 482. Kawakita, M., Yonemura, Y., Miyake, H., Ohkubo, T., Asou, N., Hayakawa, K., Nakamura, M., Kitoh, T., Osawa, H. & Takatsuki, K. Soluble c-kit molecule in serum from healthy individuals and patients with haemopoietic disorders. *British Journal of Haematology* **91**, 23–29 (1995).
- 483. Sette, C., Paronetto, M., Barchi, M., Bevilacqua, A., Geremia, R. & Rossi, P. Tr-kitinduced resumption of the cell cycle in mouse eggs requires activation of a Src-like kinase. *The EMBO Journal* 21, 5386–5395 (2002).
- 484. Mezquita, B., Mezquita, P., Pau, M., Mezquita, J. & Mezquita, C. Unlocking Doors without Keys: Activation of Src by Truncated C-terminal Intracellular Receptor Tyrosine Kinases Lacking Tyrosine Kinase Activity. *Cells* 3, 92–111 (2014).
- 485. Harpur, A., Layton, M., Das, P., Bottomley, M., Panayotou, G., Driscoll, P. & Waterfield,
  M. Intermolecular Interactions of the p85α Regulatory Subunit of Phosphatidylinositol
  3-Kinase. Journal of Biological Chemistry. 274, 12323–12332 (1999).
- 486. Fox, M., Mott, H. R. & Owen, D. Class IA PI3K regulatory subunits: p110-independent roles and structures. *Biochemical Society Transactions.* 48, 1397–1417 (2020).
- 487. Clayton, N., Fox, M., Vicente-Garcia, J., Schroeder, C., Littlewood, T., Wilde, J., Corry, J., Krishnan, K., Zhang, Q., Wakelam, M., Brown, M., Crafter, C., Mott, H. & Owen, D.

Assembly of novel, nuclear dimers of the PI3-Kinase regulatory subunits underpins the pro-proliferative activity of the Cdc42-activated tyrosine kinase, ACK. *bioRxiv.* (2019).

- 488. Park, S., Zhou, Y., Lee, J., Lu, A., Sun, C., Chung, J., Ueki, K. & Ozcan, U. The regulatory subunits of PI3K, p85α and p85β, interact with XBP-1 and increase its nuclear translocation. *Nature Medicine*. 16, 429–437 (2010).
- 489. Winnay, J., Boucher, J., Mori, M., Ueki, K. & Kahn, C. A regulatory subunit of phosphoinositide 3-kinase increases the nuclear accumulation of X-box-binding protein-1 to modulate the unfolded protein response. *Nature Medicine*. (16).
- 490. Leung, K. K., Jr Hause, R. J., Barkinge, J. L., Ciaccio, M. F., Chuu, C. P. & Jones,
  R. B. Enhanced prediction of Src homology 2 (SH2) domain binding potentials using a fluorescence polarization-derived c-Met, c-Kit, ErbB, and androgen receptor interactome.
  Mol Cell Proteomics. 13, 1705–1723 (2014).
- 491. Liu, C., Chen, T., Chau, G., Jan, Y., Chen, C., Hsu, C., Lin, K., Juang, Y., Lu, P., Cheng, H., Chen, M., Chang, C., Ting, Y., Kao, C., Hsiao, M. & Huang, C. Analysis of protein-protein interactions in cross-talk pathways reveals CRKL protein as a novel prognostic marker in hepatocellular carcinoma. *Molecular and Cellular Proteomics* 12, 1335–1349 (2013).
- 492. Rottapel, R., Reedijk, M., Williams, D., Lyman, S., Anderson, D., Pawson, T. & Bernstein, A. The Steel/W transduction pathway: kit autophosphorylation and its association with a unique subset of cytoplasmic signaling proteins is induced by the Steel factor. Molecular and Cellular Biology 11, 3043–3051 (1991).
- 493. Yi, T. & Ihle, J. Association of hematopoietic cell phosphatase with c-Kit after stimulation with c-Kit ligand. *Molecular and Cellular Biology* 13, 3350–3358 (1993).
- 494. Tauchi, T., Feng, G., Marshall, M., Shen, R., Mantel, C., Pawson, T. & Broxmeyer, H. The ubiquitously expressed Syp phosphatase interacts with c-kit and Grb2 in hematopoietic cells. *Journal of Biological Chemistry* 269, 25206–25211 (1994).
- 495. Sun, M., Hillmann, P., Hofmann, B., Hart, J. & Vogt, P. Cancer-derived mutations in the regulatory subunit p85alpha of phosphoinositide 3-kinase function through the catalytic

subunit p110alpha. Proceedings of the National Academy of Sciences of the United States of America **107**, 15547–15552 (2010).

- 496. Vallejo-Diaz, J., Chagoyen, M., Olazabal-Moran, M., Gonzalez-Garcia, A. & Carrera, A. The Opposing Roles of PIK3R1/p85a and PIK3R2/p85b in Cancer. *Trends Cancer* 5, 233–244 (2019).
- 497. Kumar, A., Redondo-Munoz, J., Perez-Garcia, V., Cortes, I., Chagoyen, M. & Carrera,
  A. Nuclear but not cytosolic phosphoinositide 3-kinase beta has an essential function in cell survival. *Molecular and Cellular Biology* 31, 2122–2133 (2011).
- 498. Hao, Y., He, B., Wu, L., Li, Y., Wang, C., Wang, T., Sun, L., Zhang, Y., Zhan, Y., Zhao, Y., Markowitz, S., Veigl, M., Conlon, R. & Wang, Z. Nuclear translocation of p85β promotes tumorigenesis of PIK3CA helical domain mutant cancer preprint on webpage at https://assets.researchsquare.com/files/rs-462020/v1\_stamped.pdf.
- 499. Olzscha, H., Schermann, S. M., Woerner, A. C., Pinkert, S., Hecht, M. H., Tartaglia, G. G., Vendruscolo, M., Hayer-Hartl, M., Hartl, F. U. & Vabulas, R. M. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell.* 144, 67–78 (2011).
- 500. Uemura, E., T, N., Minami, S., Takemoto, K., Fukuchi, S., Machida, K., Imataka, H., Ueda, T., Ota, M. & Taguchi, H. Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Scientific Reports.* 8, 678 (2018 Jan 12;).
- 501. Shigemitsu, Y. & Hiroaki, H. Common molecular pathogenesis of disease-related intrinsically disordered proteins revealed by NMR analysis. *Journal of Biochemistry.* 163, 11–18 (2018).
- Arosio, P., Vendruscolo, M., Dobson, C. M. & Knowles, T. P. Chemical kinetics for drug discovery to combat protein aggregation diseases. *Trends in Pharmacological Sciences*. 35, 127–135 (2014).
- 503. Redler, R. L., Shirvanyants, D., Dagliyan, O., Ding, F., Kim, D. N., Kota, P., Proctor,E. A., Ramachandran, S., Tandon, A. & Dokholyan, N. V. Computational approaches

to understanding protein aggregation in neurodegeneration. Journal of Molecular Cell Biology. 6, 104–115 (2014).

- 504. Yang, W., Soares, J., Greninger, P., Edelman, E., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J., Thompson, I., Ramaswamy, S., Futreal, P., Haber, D., Stratton, M., Benes, C., McDermott, U. & Garnett, M. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research.* 41, D955–D961 (2013).
- 505. Zhu, Y., Wang, Y., Guan, B., Rao, Q., Wang, J., Ma, H., Zhang, Z. & Zhou, X. C-kit and PDGFRA gene mutations in triple negative breast cancer. *International Journal of Clinical and Experimental Pathology.* 7, 4280–4285 (2014).
- 506. Finn, R., Dering, J., Ginther, C., Wilson, C., Glaspy, P., Tchekmedyian, N. & Slamon, D. Dasatinib, an orally active small molecule inhibitor of both the src and abl kinases, selectively inhibits growth of basal-type/"triple-negative" breast cancer cell lines growing in vitro. Breast Cancer Research and Treatment. 105, 319–326 (2007).
- 507. Harrell, J., Shroka, T. & Jacobsen, B. Estrogen induces c-Kit and an aggressive phenotype in a model of invasive lobular breast cancer. *Oncogenesis.* **6** (2017).
- 508. Nassar, A., Sussman, Z., Lawson, D. & Cohen, C. Inference of the Basal epithelial phenotype in breast carcinoma from differential marker expression, using tissue microarrays in triple negative breast cancer and women younger than 35. The Breast Journal. 18, 399–405 (2012).
- 509. Abbaspour Babaei, M., Kamalidehghan, B., Saleem, M., Huri, H. & Ahmadipour, F. Receptor tyrosine kinase (c-Kit) inhibitors: a potential therapeutic target in cancer cells. Drug Design, Development and Therapy 10, 2443–2459 (2016).
- 510. Ikeda, A., Judelson, D., Federman, N., Glaser, K., Landaw, E., Denny, C. & Sakamoto, K. ABT-869 inhibits the proliferation of Ewing Sarcoma cells and suppresses plateletderived growth factor receptor beta and c-KIT signaling pathways. *Molecular Cancer Therapeutics* 9, 653–660 (2010).

- 511. Wang, W., Healy, M., Sattler, M., Verma, S., Lin, J., Maulik, G., Stiles, C., Griffin, J., Johnson, B. & Salgia, R. Growth inhibition and modulation of kinase pathways of small cell lung cancer cell lines by the novel tyrosine kinase inhibitor STI 571. Oncogene. 19, 3521–3528 (2000).
- 512. Krystal, G., Honsawek, S., Kiewlich, D., Liang, C., Vasile, S., Sun, L., McMahon, G. & Lipson, K. Indolinone tyrosine kinase inhibitors block Kit activation and growth of small cell lung cancer cells. *Cancer Research* 61, 3660–3668 (2001).
- Carlino, M., Todd, J. & Rizos, H. Resistance to c-Kit inhibitors in melanoma: insights for future therapies. Oncoscience. 1, 423–426 (2014).
- 514. Iqbal, N. Imatinib: a breakthrough of targeted therapy in cancer. *Chemotherapy Research* and *Practice* (2014).
- 515. Eberle, F., Leinberger, F., Saulich, M., Seeger, W., Engenhart-Cabillic, R., J, H., K, H., Dikomey, E. & Subtil, F. In cancer cell lines inhibition of SCF/c-Kit pathway leads to radiosensitization only when SCF is strongly over-expressed. *Clin Transl Radiat Oncol.* 2, 69–75 (2017).
- 516. Elsberger, B., Fullerton, R., Zino, S., Jordan, F., Mitchell, T. J., Brunton, V. G., Mallon, E. A., Shiels, P. G. & Edwards, J. Breast cancer patients' clinical outcome measures are associated with Src kinase family member expression. *Breast Cancer Research.* 103, 899–909 (2010).
- 517. Wu, T., Wang, X., Li, J., Song, X., Wang, Y., Wang, Y., Zhang, L., Li, Z. & Tian, J. Identification of Personalized Chemoresistance Genes in Subtypes of Basal-Like Breast Cancer Based on Functional Differences Using Pathway Analysis. *PLoS One.* 10, e0131183 (2015).
- 518. Santpere, G., Alcaraz-Sanabria, A., Corrales-Sanchez, V., Pandiella, A., Gyorffy, B. & Ocana, A. Transcriptome evolution from breast epithelial cells to basal-like tumors. *Oncotarget.* 9, 453–463 (2017).
- Furth, P. A. STAT signaling in different breast cancer sub-types. Molecular and Cellular Endocrinology. 382, 612–615 (2014).
- 520. Palacios, E. H. & Weiss, A. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. Oncogene. 23, 7990–8000 (2004).
- 521. Thompson, E., Taube, J. M., Elwood, H., Sharma, R., Meeker, A., Warzecha, H. N., Argani, P., Cimino-Mathews, A. & Emens, L. A. The immune microenvironment of breast ductal carcinoma in situ. *Modern Pathology.* 29, 249–258 (2016).
- 522. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature.* **509**, 582–587 (2014).
- 523. Stein, A. & Aloy, P. Contextual specificity in peptide-mediated protein interactions. *PLoS One.* 3, e2524 (2008).
- 524. Chica, C., Diella, F. & Gibson, T. J. Evidence for the concerted evolution between short linear protein motifs and their flanking regions. *PLoS One.* **4**, e6052 (2009).
- 525. Tinti, M., Kiemer, L., Costa, S., Miller, M. L., Sacco, F., Olsen, J. V., Carducci, M., Paoluzi, S., Langone, F., Workman, C. T., Blom, N., Machida, K., Thompson, C. M., Schutkowski, M., Brunak, S., Mann, M., Mayer, B. J., Castagnoli, L. & Cesareni, G. The SH2 domain interaction landscape. *Cell Reports.* 3, 1293–305 (2013).

#### Appendix A

# Bayesian Hierarchical Clustering Algorithm

Bayesian Hierarchical Clustering (BHC) is a bottom-up hierarchical clustering method that initially treats each object as an individual cluster and then merges pairs of clusters as moving up the hierarchy, creating a tree like clustering structure. BHC uses *Dirichlet* Process to model the uncertainty in the data and uses Bayesian hypothesis testing model to decide which clusters to merge. Instead of using principled distance metrics, similarity in BHC is measured through a statistical test that evaluates marginal likelihoods of a probabilistic model. For each candidate merge, the statistical model determines and compares the probability of two hypotheses:

- all the sample in a potential merge were generated from the same mixture component;
- the sample in a potential merge came from some other clustering and have two or more clusters in it.

Assuming  $D = x^{(1)}, ..., x^{(n)}$  represents the complete sample dataset, and  $D_i \subset D$  represents the set of samples at the leaves of a subtree  $T_i$ . Initially, each sample represents a tree, with each tree containing a single sample  $D_i = x^{(i)}$ . At each stage, the algorithm considers merging all pairs of existing trees. For example, if  $T_i$  and  $T_j$  are merged into some new tree  $T_k$  then the associated set of sample is  $D_k = Di \cup Dj$  (Figure A.1 left).

Below is the code used for the BHC analysis described in this work. Example data can be found from this paper: Revealing nuclear receptor hub modules from Basal-like breast cancer

expression networks from PLOS ONE (in production, 10 June 2021). ### # Clustering Code using BHC (using BHC R/BHC and Replots) # code for reproducible research (prepared for CODECHECK codecheck. org.uk)# # This code is part of, and published under the same terms and conditions as, the following publication # Nienyun Sharon Hsu, Erika Wong En Hui, Mengzhen Liu, Di Wu, Thomas A. Hughes & James Smith, # (2021). Revealing nuclear receptor hub modules from Basal-like breast cancer expression networks. # PLOS ONE # # Code authored N Hsu and J Smith (2016-2020)# was first written for R(v3.5.3), R Studio(v1.2.5019), # R/Bioconductor (3.10), R/BHC (v1.38.0)# Last tested on R(v4.0.4), R Studio(v1.4.1103), # R/Bioconductor (3.10), R/BHC (v1.42.0)# # Recommended first use is to manually execute each Stage incrementally in RStudio. # Recommended folder organisation: # ~/ Analysis/ #~/Analysis/Data/ for .csv data input files # ~/Analysis/Working\_directory/ for Rdata files and output .txt and . pdf files# # This cluster analysis uses, # Non-parametric Multinomial Bayesian Hierarchical clustering R/ Bioconductor/BHC

```
\# for Latent feature detection in patient-gene expression data DOI:
  10.18129/B9.bioc.BHC
# R/BHC Requires a prior Bioconductor installation, see https://www.
  bioconductor.org/install/
#
# Text output files list the cluster membership.
#
\# The Visualisation requires,
# R/gplots https://cran.r-project.org/web/packages/gplots/index.
  html
\# and gplots::heatmap.2 Enhanced Heat Map to represent the
  multinomial data
\# Black = marginal likelihood, the signal contributions
# Red = upperbound, silent contributions
# Green = lowerbound, silent contributions
#
###
\#\#\# STAGE 1 - Initiation
```

```
cat("\014") ## clear RStudio console
rm(list=ls()) ## clear the Global Environment
```

#### working directory

### Example TCGA Analysis folder

setwd("~/Basal\_vs\_LuminalB/Working\_directory/")

```
### The example here is for clustering TCGA data which was already median-centred.
```

 $\#\!/\!/\!/\!/$  For the BRCA\_METABRIC data, the alternative code was employed

```
### TCGA
\#\#\# Example .csv file, see SI_3_4 Original data was median-centred
    (autoscaled)
signal_data <- read.csv("../Data/TCGA_BA+LUB_171genes.csv", row.
   names = 1)
### METABRIC
\#\#\# load .csv file
# metabric_data <- read.csv("../Data/METABRIC_BA+LUB_169genes.csv",
   row.names = 1, col.names = ,1)
# metabric_data <- metabric_data[-1,] # optional
\#\#\# median-centre
# rowmed <- apply(metabric_data,1,median)</pre>
# mediancentred_data <- metabric_data - rowmed</pre>
# signal_data <- mediancentred_data
textRowLabel <--"Genes"
textColLabel <--"Patients"</pre>
save(textRowLabel, file="textRowLabel.Rda")
save(textColLabel, file="textColLabel.Rda")
#### Specific row and col editing for TCGA data
n_col<-ncol(signal_data)
n_row<-nrow(signal_data)
new_row_names <- rep("X",n_row)
for (i in seq(n_row)){ new_row_names[i]<-paste0(row.names(signal_
   data) [i]) }
row.names(signal_data) <- new_row_names</pre>
#### Example for TCGA data here
```

```
for (i in seq(n_col)){ colnames(signal_data)[i] <-gsub("TCGA.", "",
    colnames(signal_data)[i]) }
```

X\_original <- as.data.frame(signal\_data)

```
### STAGE 2 - Repeat this for re-calibration of thresholds, if
   necessary.
\#\#\# information content assessment is row and col elimination if too
    many 0s \ (== N/As!)
#### examination of data with large fraction of missing values
   requires calibration
###
### Subset_data only keeps rows and columns with complete
   information below threshold
#### (only minimal amounts of missing data)
### Replace NAs with 0
X <- X_original
X[is.na(X)] <- 0
\#//// row filter – acceptable threshold of missing data in rows
### row threshold established empirically for TCGA and METABRIC data
\#\#\# eg 1/8 is the most generous, 1/16, 1/32, 1/64 etc
threshold_r <- 1/64
row_data_flag \leftarrow rep(1, nrow(X))
for (i \text{ in } seq(nrow(X))) \{ if(length(which(X[i,]==0))) > threshold_r*(
   ncol(X))) { row_data_flag[i] < -0 }
acceptable_rows <- which (row_data_flag==1) ### important
rejected_rows <- which (row_data_flag==0) ### important information
```

length(rejected\_rows) #### for information

rejectedRData\_file <- paste0("rejectedrows", ".Rda")</pre>

```
save(rejected_rows, file=rejectedRData_file) #### saved for
information
```

### column filter – acceptable threshold of missing data in rows #### col threshold established empirically for TCGA and METABRIC data #### eg 1/8 is the most generous, 1/16, 1/32, 1/64 etc threshold\_c <-1/32

 $col_data_flag \leftarrow rep(1, ncol(X))$ 

for (i in seq(ncol(X))){ if (length(which(X[,i]==0)) > threshold\_c\*(
 nrow(X))) { col\_data\_flag[i]<-0} }
acceptable\_cols <- which(col\_data\_flag==1) ### important
rejected\_cols <- colnames(X)[which(col\_data\_flag==0)]
length(rejected\_cols) ### for information
rejectedCData\_file <- paste0("rejectedcols", ".Rda")
save(rejected\_cols, file=rejectedCData\_file) ### saved for
 information</pre>

```
subset_data <- X[acceptable_rows, acceptable_cols] #### subsetting for
Complete data
```

### STAGE 3 (optional)

#### To show ordination does not bias the process, we randomise rows
and columns.

### We believe BHC is not affected by order.

#### Warning heatmap.2 function visualisation compositing might break
down with v large data,

 $\#\!\#\!\#\!$  so user could test with a fraction of the data first:

 $data_fraction \ll 1.0 \# \# 1.0 = maximum$ 

272

```
new_r_length <- as.integer(data_fraction*length(acceptable_rows))
new_c_length <- as.integer(data_fraction*length(acceptable_cols))</pre>
```

```
### Randomise rows and columns
rnd_pointers_r <- sample(acceptable_rows, new_r_length, replace=FALSE
)
rnd_pointers_c <- sample(acceptable_cols, new_c_length, replace=FALSE
)</pre>
```

```
\#\!/\!\!/\!\!/\!\!/ subsetting for Complete data
```

```
subset_data <- as.data.frame(matrix(0, nrow=new_r_length , ncol=new_
c_length))
for(i in seq(1:length(rnd_pointers_r))) {
    for(j in seq(1:length(rnd_pointers_c))) {
        subset_data[i,j] <- X[rnd_pointers_r[i], rnd_pointers_c[j]]
        }
    }
    row.names(subset_data) <- rownames(X)[rnd_pointers_r]
    colnames(subset_data) <- colnames(X)[rnd_pointers_c]</pre>
```

**rm**(i,j)

### doing this to save memory
rm(signal\_data,X) #### optional

Xbhc <- subset\_data
rm(subset\_data) ### optional
save (Xbhc, file="Xbhc.Rda") ### backup</pre>

#### library (BHC)

```
### Based on R/BHC Examples - FOR THE MULTINOMIAL CASE, THE DATA CAN
BE DISCRETISED
col_names_for_BHC <- colnames(Xbhc)
row_names_for_BHC <- row.names(Xbhc)
#### Assume n > p
```

nDataItems < **nrow**(Xbhc) ### (n) nFeatures < **ncol**(Xbhc) ### (p)

```
### numThreads is 1 (default) or more, depending on the available
    cores for speed
```

nt <- 2

```
### robustFlag is 0 (default) to use single Gaussian likelihood, 1
to use mixture likelihood.
```

robust <- 1

- percentiles\_a <- FindOptimalBinning(Xbhc, row\_names\_for\_BHC, transposeData=TRUE, verbose=TRUE)
- discreteData\_temp <- DiscretiseData(t(Xbhc), percentiles=percentiles \_a) # apply to transform by default
- discreteData\_a <- t(discreteData\_temp)

save(discreteData\_a, file="discretizedData\_a.Rda")

```
save(percentiles_a, file="percentiles_a.Rda")
```

- **rm**(discreteData\_temp)
- hc\_a <- bhc(discreteData\_a, row\_names\_for\_BHC, verbose=TRUE, robust= robustFlag, numThreads=nt)

```
save(hc_a, file = "hc_a.Rda")
```

```
txt_file_a <- paste0("hc_labels_",textRowLabel,"_a.txt")</pre>
```

WriteOutClusterLabels(hc\_a, txt\_file\_a, verbose=TRUE)

- ### requires transform of the data as t(Xbhc), be mindful of the
  interpretation
- percentiles\_b <- FindOptimalBinning(t(Xbhc), col\_names\_for\_BHC,

transposeData=TRUE, verbose=TRUE)

discreteData\_temp <- DiscretiseData(Xbhc, percentiles=percentiles\_b)

# apply to transform by default

discreteData\_b <- t(discreteData\_temp)

save(discreteData\_b, file="discretizedData\_b.Rda")

save(percentiles\_b, file="percentiles\_b.Rda")

**rm**(discreteData\_temp)

hc\_b <- bhc(discreteData\_b, col\_names\_for\_BHC, verbose=TRUE, robust= robustFlag, numThreads=nt)

 $save(hc_b, file = "hc_b.Rda")$ 

txt\_file\_b <- paste0("hc\_labels\_",textColLabel,"\_b.txt")</pre>

WriteOutClusterLabels(hc\_b, txt\_file\_b, verbose=TRUE)

library (gplots)

cat("\014") ### clear RStudio console
rm(list=ls()) ### clear the Global Environment

# Example TCGA Analysis folder
setwd("~/Basal\_vs\_LuminalB/Working\_directory/")
#### Load data
load("Xbhc.Rda")
load("hc\_a.Rda")

- load ("hc\_b.Rda")
- **load** ("discretized Data\_a.Rda")
- **load** ("discretized Data\_b.Rda")
- load("textColLabel.Rda")
- load("textRowLabel.Rda")
- ### Heatmap with heatmap.2() for .pdf file export using Rstudio
  #### Remember lower bound = green, Marginal Likelihood = black!,
  upper bound = red
- ### Important Note: With newer versions of BHC and gplots libraries, #### the row and column labels are NOT aligned with the plot row widths and column

### widths in the heatmaps. It is important therefore to refer to
#### hc\_labels\_a.txt (see above) and hc\_labels\_b.txt (see above).

- $col_percentiles <- c( ((100-ML_a)/2)*0.01,ML_a*0.01,((100-ML_a)/2)*0.01)$
- text\_col\_percent <- paste(as.character(col\_percentiles), collapse=", \_")
- table\_text <- paste0(nrow(Xbhc),"\_",textRowLabel,"\_by\_",ncol(Xbhc),"
  \_", textColLabel,"\_(",text\_col\_percent,")")</pre>

heatmap.2(t(discreteData\_a),trace="none", col=greenred(256),

dendrogram="both", key=FALSE, Rowv=hc\_b, Colv=hc\_a)

dev.copy(pdf,paste0("Rplot\_",table\_text,"\_a.pdf"))

```
\mathbf{dev}. off()
```

0.01)

```
text_row_percent <- paste(as.character(row_percentiles), collapse=",
_")
```

table\_text <- paste0(ncol(Xbhc),"\_",textColLabel,"\_by\_",nrow(Xbhc),"
\_", textRowLabel, "\_(",text\_row\_percent,")")</pre>

heatmap.2(t(discreteData\_b),trace="none", col=greenred(256),

dendrogram="both", key=FALSE, Rowv=hc\_a, Colv=hc\_b)

**dev**.copy(pdf,paste0("Rplot\_",**table\_text**,"\_b.pdf"))

 ${\bf dev}\,.\,{\bf off}\,(\,)$ 

 $\# \ end$ 



Figure A.1: (Left) Schematic of a portion of a tree where  $T_i$  and  $T_j$  are merged into  $T_k$ , and the associated sample sets  $D_i$  and  $D_j$  are merged into  $D_k$ . (Right) An example tree with samples. The clusterings  $(1 \ 2)(3)(4)$  or  $(1 \ 2 \ 3)(4)$  are tree-consistent partitions whereas the clustering (1)(23)(4) or  $(1 \ 3)(2 \ 4)$  are not a tree-consistent partitions.

In order to determine optimal merges, BHC algorithm compares two hypotheses. The first hypothesis  $(H_1^k)$  assumes that all the samples in  $D_k$  were generated independently and identically from the same probabilistic model,  $p(x|\theta)$  with unknown parameters  $\theta$ . Assuming that the probabilistic model is a multivariate Gaussian with parameters  $\theta = (\mu, \Sigma)$ , the marginal likelihood of the sample under this hypothesis can be computed using Equation A.1, where  $p(\theta|\beta)$  is the prior over the parameters of the model, with hyperparameters  $\beta$ .

$$p(D_k|H_1^k) = \int p(D_k|\theta)p(\theta|\beta)d\theta$$
$$= \int \left[\prod_{x^{(i)} \in D_k} p(x^{(i)}|\theta)\right] p(\theta|\beta)d\theta$$
(A.1)

This equation calculates the probability that all the samples in  $D_k$  were generated from the same parameter values assuming a model of the form  $p(x|\theta)$ , representing a natural model-based criterion for measuring how well the samples fit into one cluster.

The second hypothesis  $(H_2^k)$  assumes that samples in  $D_k$  are form more than one cluster. While summing the possible ways of dividing  $D_k$  into more than one cluster is intractable, the sum can be computed efficiently by recursion if clusterings are restricted to only tree-consistent partitions (Figure A.1 right). The probability of the sample under this restricted alternative hypothesis, is simply a product over the subtrees  $p(D_k|H_2^k) = p(D_i|T_i)p(D_j|T_j)$  where  $p(D_i|T_i)$ is the probability of a sample set under a tree.

Combining the probability of the sample under the two hypotheses,  $H_1^k$  and  $H_2^k$ , weighted by the prior that all samples in  $D_k$  belong to one cluster,  $\pi_k \stackrel{\text{def}}{=} p(H_1^k)$ . The marginal probability of the sample in tree  $T_k$  can be computed using the following equation:

$$p(D_k|T_k) = \pi_k p(D_k|H_1^k) + (1 - \pi_k) p(D_i|T_i) p(D_j|T_j)$$
(A.2)

Equation A.2 is defined recursively, where the first term considers the hypothesis that there is a single cluster in  $D_k$  and the second term sums over all other possible tree-consistent partitions in  $D_k$  (Figure A.1). This equation can be used to derive an approximation to the marginal likelihood of a *Dirichlet* Process mixture model, providing a new lower bound on this marginal likelihood. Importantly, the prior for the merged hypothesis,  $\pi_k$ , can be computed bottom-up in a *Dirichlet* Process mixture and the posterior probability of the merged hypothesis,  $r_k \stackrel{\text{def}}{=} p(H_1^k | D_k)$ , can obtained using the Bayes rule:

$$r_k = \frac{\pi_k p(D_k | H_1^k)}{\pi_k p(D_k | H_1^k) + (1 - \pi_k) p(D_i | T_i) p(D_j | T_j)}$$
(A.3)

The quantity calculated by Equation A.3 is used to decide which two trees to merge and which merges the algorithm would prefer not to make given the priors. The general BHC algorithm is described in Figure A.2. The context, equations and figures described in this appendix were adapted from Heller [182], and more details regarding the BHC algorithm and its application can be found in [183, 184, 186].

input: data  $\mathcal{D} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(n)}\}$ , model  $p(\mathbf{x}|\theta)$ , prior  $p(\theta|\beta)$ initialize: number of clusters c = n, and  $\mathcal{D}_i = \{\mathbf{x}^{(i)}\}$  for  $i = 1 \dots n$ while c > 1 do Find the pair  $\mathcal{D}_i$  and  $\mathcal{D}_j$  with the highest probability of the merged hypothesis:  $r_k = \frac{\pi_k p(\mathcal{D}_k | \mathcal{H}_1^k)}{p(\mathcal{D}_k | T_k)}$ Merge  $\mathcal{D}_k \leftarrow \mathcal{D}_i \cup \mathcal{D}_j, \quad T_k \leftarrow (T_i, T_j)$ Delete  $D_i$  and  $D_j, c \leftarrow c - 1$ end while output: Bayesian mixture model where each tree node is a mixture component The tree can be cut at points where  $r_k < 0.5$ 

Figure A.2: Bayesian hierarchical clustering algorithm.

Appendix B

Nuclear Receptors and Coregulators

Table B.1: The 48 human nuclear receptors, with their corresponding ligands and coregulators. There are a total number of 178 non-redundant genes, including the 48 nuclear receptors and coregulators.

Name	NRNC	Gene	Ligands	Coregulators
	Symbol	Names		
Thyroid hormone receptor- $\alpha$	NR1A1	THRA		ESR1, MED1, MED24, NCOA1, NCOA6,
			тиугою погшове	NCOR1, NCOR2, NR0B2, NR2F1,
				NRIP1
Thyroid hormone receptor- $\beta$	NR1A2	THRB		ESR1, MED1, NCOA1, NCOA2, NCOA3,
				NCOA6, NCOR1, NCOR2, RXRA,
				RXRG
Retinoic acid receptor- $\alpha$	NR1B1	RARA		RXRG, PML, NRIP1, NCOR2, NCOR1,
			Vitamin A	NCOA3, NCOA2, NCOA1, MED1,
				KAT2B
Retinoic acid receptor- $\beta$	NR1B2	RARB		ESR2, HNF4G, NR0B2, NR2C1, NR2F1,
				NR4A1, NR4A2, RARA, RORC, RXRG
Retinoic acid receptor- $\gamma$	NR1B3	RARG		HNF4G, MED1, NR0B2, NR2F6,
				NR4A1, NR4A2, NRBP1, RORB,
				RXRG, VDR

Peroxisome prolifer	ator-	NR1C1	PPARA		ABCA1, EP300, FABP1, HSP90AA1,
activated receptor- $\alpha$				Fatty acids, prostaglandins	LPL, MED1, NCOA1, NCOR1, NCOR2,
					PPARGC1A
Peroxisome prolifer	ator-	NR1C2	PPARD		BCL6, CREBBP, CTNNB1, EP300,
activated receptor- $\beta/\delta$					HNF4A, MED1, NR0B2, NR4A1,
					NR4A2, RXRG
Peroxisome prolifer	ator-	NR1C3	PPARG	<u>.</u>	PPARGC1A, NCOR2, NCOR1, NCOA2,
activated receptor- $\gamma$					NCOA1, MED1, LEP, HDAC3, EP300,
					ADIPOQ
${ m Rev-ErbA}lpha$		NR1D1	NR1D1	· · · · · 11	ARNTL, CLOCK, EP300, HDAC3,
				пеще	NCOR1, NPAS2, NR0B2, NR1D2,
					RORA, RORC
${ m Rev-ErbA}lpha$		NR1D2	NR1D2	<u>.</u>	ARNTL, ESR2, NCOR1, NR0B2,
					NR1D1, NR4A1, RORA, RORB, RORC,
					SKP1
RAR-related or	phan	NR1F1	RORA		ARNTL, CLOCK, EP300, MED1,
receptor- $\alpha$				Cholesterol, ATRA	NCOA1, NCOA2, NPAS2, NR1D2,
					SMARCD3, VANGL2

RAR-related orph	an NR1F2	RORB		NR0B2, NR1D1, NR1D2, NR1H2,
$\operatorname{receptor}-eta$				NR2C1, NR4A2, NR5A1, NRBP1,
				RORC, RXRG
RAR-related orph	an NR1F3	RORC		FOXP3, NPAS2, NR0B2, NR1D1,
receptor- $\gamma$				NR1D2, NR1H2, NR2C1, NR5A1,
				RORB, RXRG
Liver X receptor- $\beta$	NR1H2	NR1H2		ABCA1, ABCG8, ESR1, NR0B2,
			Oxysterols	NR2F1, NR4A1, PPARA, PPARG,
				RXRA, RXRG
Liver X receptor- $\alpha$	NR1H3	NR1H3		ABCA1, APOA1, LPL, NCOA1, NR0B2,
				NRIP1, PPARA, PPARGC1A, RXRB,
				RXRG
Farnesoid X receptor	NR1H4	NR1H4		ABCB11, CYP7A1, EIF2C2, FABP6,
				GPS2, INS, MED1, NISCH, NR0B2,
				RXRG
Vitamin D receptor	NR111	VDR	vitamin D	TCF3, SRC, SMAD3, NCOR1, NCOA3,
				NCOA2, NCOA1, MED1, EP300,
				CREBBP

Pregnane X receptor	NR1I2	NR112	xenobiotics	ABCB1, ABCB11, CYP2B6, CYP3A4,
				CYP7A1, NCOA1, NCOR1, NCOR2,
				RXRA, SLC01B1
Constitutive androstane re-	NR1I3	NR113	androstane	USF2, THRSP, RXRA, NR0B1,
ceptor				MED1, HSP90AA1, CYP3A4, CYP2E1,
				CYP2B6, CYP1A2
Hepatocyte nuclear factor-4-	NR2A1	HNF4A	To 4444	SMAD4, SMAD3, PPARGC1A, NR0B2,
α			rauty actus	MED1, HNF1A, FOX01, CTNNB1,
				CREBBP, APOA1
Hepatocyte nuclear factor-4-	NR2A2	HNF4G		THRA, NRBP1, NR2C2, NR2C1,
7				NR1H2, NR0B2, MED1, HNF1A, ES-
				RRG, ESRRA
Retinoid X receptor- $\alpha$	NR2B1	RXRA		NR1H2, NR1H3, NR1I2, NR1I3, PPARG,
			Retinoids	RARA, RARS, STAT1, THRB, VDR
Retinoid X receptor- $\beta$	NR2B2	RXRB		FKBP1A, FKBP1B, NCOA3, NR1H2,
				NR1H3, NR4A2, RARB, RARS, RGL2,
				UBE2I
Retinoid X receptor- $\gamma$	NR2B3	RXRG		NR1H2, NR1H3, NR4A1, PPARG,
				RARA, RARB, RARG, THRA, THRB,
				VDR

Testicular receptor 2	NR2C1	NR2C1	II. I	NRIP1, NR5A1, NR4A2, NR4A1,
			ошкломи (огрнан гесериог)	NR2F6, NR2C2, NR1D1, HDAC3, ESR1,
				AR
Testicular receptor 4	NR2C2	NR2C2		AR, ESR1, HNF4A, IKBKB, JAZF1,
				NR2C1, NR2C2AP, NR4A2, NR5A1,
				TAB2
Homologue of the Drosophila	NR2E1	NR2E1	II	RORB, RERE, NR5A1, NR4A2, NR4A1,
tailless gene			онкломи (огрнал гесериог)	NR2C2, NR2C1, NR1D1, HNF4G,
				HNF4A
Photoreceptor cell-specific	NR2E3	NR2E3		CRX, NRL, RHO, RPE65, VSX2,
nuclear receptor				PDE6B, PRPH2, OPN1SW, OTX2,
				ENSG00000235718
Chicken ovalbumin upstream	NR2F1	NR2F1		THRA, RXRG, RERE, NR4A2, NR4A1,
promoter-transcription factor			Retinoic acid	NR2C1, NR1H2, MED1, HNF4A, ESR1
I				
Chicken ovalbumin upstream	NR2F2	NR2F2		APOB, BCL11A, DR1, KLF4, LCK,
promoter-transcription factor				NRP2, PAX6, PCK2, RERE, SOX9
П				
V-erbA-related	NR2F6	NR2F6		ESR1, NR1H2, NR2C1, NR4A1, NR4A2,
				NRBP1, RARG, RASD1, RXRG, THRB

Estrogen receptor- $\alpha$	NR3A1	ESR1		BRCA1, CREBBP, EP300, HDAC1,
			SUDGETTS	NCOA1, NCOA2, NCOA3, NRIP1, SP1,
				SRC
Estrogen receptor- $\beta$	NR3A2	ESR2		ESR1, FOS, MED1, NCOA1, NCOA2,
				NCOA3, NCOR2, NOS3, NR0B1, SRC
Estrogen-related receptor- $\alpha$	NR3B1	ESRRA		SIRT1, PPARGC1B, PPARGC1A,
			Unknown (orphan receptor)	NRIP1, NRF1, NR0B2, NCOR1, ES-
				RRG, ESR1, CREB1
Estrogen-related receptor- $\beta$	NR3B2	ESRRB		HNF4A, HNF4G, NANOG, NCOA3,
				NR0B2, NR1D1, NR4A1, NR4A2,
				NRBP1
Estrogen-related receptor- $\gamma$	NR3B3	ESRRG		ESRRA, HNF4G, MED1, NR0B2,
				NRBP1, PNRC2, PPARG, PPARGC1A,
				PPARGC1B, RORB
Glucocorticoid receptor	NR3C1	NR3C1	cortisol	UBC, SMARCA4, RELA, NFKB1,
				NCOA3, NCOA2, NCOA1, JUN,
				HSP90AA1
Mineralocorticoid receptor	NR3C2	NR3C2	aldosterone	FKBP4, HSP90AA1, NRIP1, RPS27A,
				SCNN1A, SGK1, SRC, SUMO1,
				TRIM24, UBC

Progesterone receptor	NR3C3	PGR	progesterone	CCND1, CDK2, ERBB2, ESR1, MAPK1,
				NCOA3, PGR, SP1, SRC
Androgen receptor	NR3C4	AR	testosterone	CTNNB1, HSP90AA1, KLK3, NCOA1,
				NCOA2, NCOA3, NCOR1, RNF14, SRC,
				TMPRSS2
Nerve Growth factor IB	NR4A1	NR4A1		AKT1, BCL2, EP300, MAPK8, MED1,
			Unknown (orphan receptor)	MEF2D, NCOR2 NR0B2, POMC, RXRG
Nuclear receptor related 1	NR4A2	NR4A2		CREB1, LMX1A, LMX1B, NR0B2,
				NR2C1, PITX3, RARA, RARB, RXRG
Neuron-derived orphan re-	NR4A3	NR4A3		EWSR1, GCH1, HIVEP1, PSMC1,
ceptor 1				PSMD2, SIX3, STX17, TAF15, TCF12,
				TFG
Steroidogenic factor 1	NR5A1	NR5A1	phosphatidylinositols	AR, CYP19A1, FOXL2, NR0B1, NR0B2,
				NRIP1, POU5F1, SOX2, SOX9, TRERF1
Liver receptor homolog-1	NR5A2	NR5A2	phosphatidylinositols	NCOA2, NCOA3, GREB1, ESR1,
				NR0B2, EP300, NR0B1, CHD1,
				HIST1H2BA, CREBBP

Germ cell nuclear factor	NR6A1	NR6A1	unknown (orphan receptor)	MUL1, ATP6V0A2, UIMC1, POU5F1,
				NANOG, ERCC6L, NCOR1, DNMT3A,
				VRTN, RMI1
Dosage-sensitive sex reversal,	NR0B1	NR0B1	Dosage-sensitive sex reversal,	AR, COPS2, CYP19A1, ESR1, ESR2,
adrenal hypoplasia critical re-			adrenal hypoplasia critical re-	NANOG, NR5A1, NRIP1, STAR, WT1
gion, on chromosome X, gene			gion, on chromosome X, gene	
1			1	
Small heterodimer partner	NR0B2	NR0B2	Small heterodimer partner	ESR1, ESRRG, HDAC1, HNF4A,
				NR1H2, NR4A1, NR5A1, NR5A2,
				RARA, SMARCA2

#### Appendix C

#### Nuclear Receptor Availability

This supplementary table summarises the gene expression available in TCGA and METABRIC dataset. TCGA analysis considered 171 genes and is missing: AGO2, KAT2B, NR2C2AP, SGK1, SKP1, TAB2, VANGL2. METABRIC analysis considered 169 genes and is missing: ADIPOQ, CYP1A2, CYP2B6, AGO2, HNF4G, INS. NANOG, SCNN1A, TCF3.

	TCGA	METABRIC
ABCA1	Present	Present
ABCB1	Present	Present
ABCB11	Present	Present
ABCG8	Present	Present
ADIPOQ	Present	Absent
AGO2	Absent	Absent
AKT1	Present	Present
APOA1	Present	Present
АРОВ	Present	Present
AR	Present	Present
ARNTL	Present	Present
BCL11A	Present	Present
BCL2	Present	Present
BCL6	Present	Present
BRCA1	Present	Present
CCND1	Present	Present

	TCGA	METABRIC
CDK2	Present	Present
CLOCK	Present	Present
COPS2	Present	Present
CREB1	Present	Present
CREBBP	Present	Present
CTNNB1	Present	Present
CYP19A1	Present	Present
CYP1A2	Present	Absent
CYP2B6	Present	Absent
CYP2E1	Present	Present
CYP3A4	Present	Present
CYP7A1	Present	Present
DR1	Present	Present
EP300	Present	Present
ERBB2	Present	Present
ESR1	Present	Present
ESR2	Present	Present
ESRRA	Present	Present
ESRRB	Present	Present
ESRRG	Present	Present
EWSR1	Present	Present
FABP1	Present	Present
FABP6	Present	Present
FKBP1A	Present	Present
FKBP1B	Present	Present
FKBP4	Present	Present
FOS	Present	Present
FOXL2	Present	Present
FOXO1	Present	Present
FOXP3	Present	Present
GCH1	Present	Present

	TCGA	METABRIC
GPS2	Present	Present
HDAC1	Present	Present
HDAC3	Present	Present
HIVEP1	Present	Present
HNF1A	Present	Present
HNF4A	Present	Present
HNF4G	Present	Absent
HSP90AA1	Present	Present
IKBKB	Present	Present
INS	Present	Absent
JAZF1	Present	Present
JUN	Present	Present
KAT2B	Absent	Present
KLF4	Present	Present
KLK3	Present	Present
LCK	Present	Present
LEP	Present	Present
LMX1A	Present	Present
LMX1B	Present	Present
LPL	Present	Present
MAPK1	Present	Present
MAPK8	Present	Present
MED1	Present	Present
MED24	Present	Present
MEF2D	Present	Present
NANOG	Present	Absent
NCOA1	Present	Present
NCOA2	Present	Present
NCOA3	Present	Present
NCOA6	Present	Present
NCOR1	Present	Present

	TCGA	METABRIC
NCOR2	Present	Present
NFKB1	Present	Present
NISCH	Present	Present
NOS3	Present	Present
NPAS2	Present	Present
NR0B1	Present	Present
NR0B2	Present	Present
NR1D1	Present	Present
NR1D2	Present	Present
NR1H2	Present	Present
NR1H3	Present	Present
NR1H4	Present	Present
NR1I2	Present	Present
NR1I3	Present	Present
NR2C1	Present	Present
NR2C2	Present	Present
NR2C2AP	Absent	Present
NR2E1	Present	Present
NR2E3	Present	Present
NR2F1	Present	Present
NR2F2	Present	Present
NR2F6	Present	Present
NR3C1	Present	Present
NR3C2	Present	Present
NR4A1	Present	Present
NR4A2	Present	Present
NR4A3	Present	Present
NR5A1	Present	Present
NR5A2	Present	Present
NRBP1	Present	Present
NRF1	Present	Present

	TCGA	METABRIC
NRIP1	Present	Present
NRP2	Present	Present
PAX6	Present	Present
PCK2	Present	Present
PGR	Present	Present
PITX3	Present	Present
PML	Present	Present
PNRC2	Present	Present
POMC	Present	Present
POU5F1	Present	Present
PPARA	Present	Present
PPARD	Present	Present
PPARG	Present	Present
PPARGC1A	Present	Present
PPARGC1B	Present	Present
PSMC1	Present	Present
PSMD2	Present	Present
RARA	Present	Present
RARB	Present	Present
RARG	Present	Present
RARS	Present	Present
RASD1	Present	Present
RELA	Present	Present
RERE	Present	Present
RGL2	Present	Present
RNF14	Present	Present
RORA	Present	Present
RORB	Present	Present
RORC	Present	Present
RPS27A	Present	Present
RXRA	Present	Present

	TCGA	METABRIC
RXRB	Present	Present
RXRG	Present	Present
SCNN1A	Present	Absent
SGK1	Absent	Present
SIRT1	Present	Present
SIX3	Present	Present
SKP1	Absent	Present
SLCO1B1	Present	Present
SMAD3	Present	Present
SMAD4	Present	Present
SMARCA2	Present	Present
SMARCA4	Present	Present
SMARCD3	Present	Present
SOX2	Present	Present
SOX9	Present	Present
SP1	Present	Present
SRC	Present	Present
STAR	Present	Present
STAT1	Present	Present
STX17	Present	Present
SUMO1	Present	Present
TAB2	Absent	Present
TAF15	Present	Present
TCF12	Present	Present
TCF3	Present	Absent
TFG	Present	Present
THRA	Present	Present
THRB	Present	Present
THRSP	Present	Present
TMPRSS2	Present	Present
TRERF1	Present	Present

	TCGA	METABRIC
TRIM24	Present	Present
UBC	Present	Present
UBE2I	Present	Present
USF2	Present	Present
VANGL2	Absent	Present
VDR	Present	Present
WT1	Present	Present

#### Appendix D

### **Patient Clustering Consolidation**



Figure D.1: Clusters to classes simplification from the Basal vs Luminal A analyses.



Figure D.2: Clusters to classes simplification from the Basal vs Luminal B analyses.



Figure D.3: Clusters to classes simplification from the Basal vs Her2 analyses.

#### Appendix E

## **Partial Correlation Networks**



Figure E.1: The partial correlation network derived from TCGA class 1 from the Basal vs Luminal A analysis.



Figure E.2: The partial correlation network derived from METABRIC class 1 from the Basal vs Luminal A analysis.



Figure E.3: The partial correlation network derived from TCGA class 1 from the Basal vs Luminal B analysis.


Figure E.4: The partial correlation network derived from METABRIC class 1 from the Basal vs Luminal B analysis.



Figure E.5: The partial correlation network derived from METABRIC class 2 from the Basal vs Luminal B analysis.



Figure E.6: The partial correlation network derived from TCGA class 1 from the Basal vs Her2 analysis.



Figure E.7: The partial correlation network derived from TCGA class 3 from the Basal vs Her2 analysis.



Figure E.8: The partial correlation network derived from METABRIC class 2 from the Basal vs Her2 analysis.

#### Appendix F

#### Nodes and Degree

This supplementary table summarises the degree (number of connections) that each node makes in each Basal-specific network. The nodes are ranked according to the magnitude of the total degree across networks. FOS and STAT1 are the most highly connected nodes, and are defined as hubs in this work. **BLA T1**: Basal vs Luminal A TCGA class 1; **BLA M1**: Basal vs Luminal A METABRIC class 1; **BLB T1**: Basal vs Luminal B TCGA class 1; **BLB M1**: Basal vs Luminal B METABRIC class 1; **BLB M2**: Basal vs Luminal B METABRIC class 2; **BH T1**: Basal vs Her2 TCGA class 1; **BH T3**: Basal vs Her2 TCGA class 3; **BH M2**: Basal vs Her2 METABRIC class 2.

	BLA T1	BLA M1	BLB T1	BLB M1	BLB M2	BH T1	BH T3	BH M2	Total
FOS	3	4	3	3	5	5	1	4	28
STAT1	4	4	2	3	4	3	3	3	26
NR2C1	3	2	3	5	3	2	4	2	24
NR4A2	4	4	3	0	3	3	3	4	24
CREBBP	3	5	4	1	4	2	0	4	23
NCOR2	3	3	3	3	3	2	3	2	22
APOA1	3	2	4	1	1	3	4	3	21
LCK	4	3	3	1	3	2	3	2	21
TMPRSS2	3	4	3	3	2	2	1	3	21
POMC	1	3	1	4	2	3	4	2	20
SP1	3	1	5	3	2	1	4	1	20
THRSP	3	3	2	2	1	2	5	2	20
USF2	3	1	2	4	0	2	5	3	20
COPS2	0	4	2	3	4	2	2	2	19
HNF4A	3	3	2	2	3	2	1	3	19
JUN	3	2	4	0	3	2	2	3	19
NR4A1	3	5	2	2	2	1	1	3	19
EWSR1	3	3	2	1	3	1	3	2	18
HSP90AA1	3	2	2	0	3	1	4	3	18
LEP	2	2	3	3	1	1	3	3	18
NR0B2	4	1	4	2	0	1	4	2	18
NR1H3	2	3	2	2	2	2	2	3	18
NR2F6	3	2	2	2	1	3	2	3	18
NR5A1	2	2	4	3	1	4	1	1	18

	BLA T1	BLA M1	BLB T1	BLB M1	BLB M2	BH T1	вн тз	BH M2	Total
PGR	3	3	3	2	2	4	0	1	18
SRC	4	3	3	2	2	3	1	0	18
STAR	2	2	4	3	2	2	1	2	18
TRIM24	2	2	2	3	1	2	5	1	18
APOB	5	0	5	2	1	2	1	1	17
LPL	1	2	3	2	1	3	3	2	17
NR3C2	2	1	1	5	2	1	3	2	17
NRP2	1	3	0	3	2	4	2	2	17
SIRT1	1	2	2	4	2	1	2	3	17
SMARCD3	1	2	2	3	3	1	2	3	17
SUMO1	1	3	3	2	3	2	2	1	17
TAF15	3	3	2	2	2	1	1	3	17
UBC	1	2	1	2	3	1	3	4	17
BRCA1	3	1	2	0	1	2	5	2	16
CYP3A4	3	0	3	2	1	3	3	1	16
ERBB2	1	1	3	2	2	2	4	1	16
FABP1	5	1	4	1	1	3	0	1	16
KLK3	1	2	3	2	3	3	1	1	16
PML	1	2	2	1	2	3	2	3	16
PPARG	2	2	7	0	1	1	2	1	16
RNF14	2	3	2	0	3	1	2	3	16
SIX3	2	2	2	2	2	2	3	1	16
THRA	1	4	0	2	6	0	0	3	16
ABCB1	2	2	1	3	3	1	1	2	15
CYP7A1	2	3	3	2	2	0	2	1	15
FABP6	1	2	0	3	2	3	1	3	15
FOXP3	2	2	2	2	2	1	1	3	15
GPS2	2	1	5	2	3	0	0	2	15
IKBKB	2	0	1	2	2	4	3	1	15
MED24	1	2	1	3	0	4	3	1	15
NR1H4	1	1	2	2	1	3	2	3	15
NR2F2	3	1	1	3	2	2	1	2	15
NR5A2	3	1	3	4	1	3	0	0	15
PAX6	1	3	0	3	3	1	2	2	15
PNRC2	2	2	3	2	2	2	0	2	15
PPARGC1A	2	3	1	1	4	0	2	2	15
PSMC1	2	3	2	0	4	1	0	3	15
RASD1	1	3	0	2	1	2	2	4	15
RERE	3	1	2	4	1	3	0	1	15
RORA	2		1	5		1	2	2	15
TUBB	2	2	2	3		1		4	15
IDEN	1	1	2	3 9	1	1	4	1	15
WT1	2	2	2	1	1	1	4	2	15
ADIPOO	6	2	3	0		1	1	0	14
ESBRA	1		3	2		4	1	2	14
MAPKS	2	4	2	2	1	3	2	1	14
NCOA2	1	3	1	0	5	1	1	2	14
NR113	1	1	2	3	1	3	2	-	14
NR4A3	2	2	4	0	0	3	1	2	14
PPARGC1B	1	3	0	1	1	4	1	3	14
RARS	1	2	1	1	3	3	2	1	14
SLCO1B1	2	2	2	3	1	3	0	1	14
SMARCA4	1	2	1	2	1	2	3	2	14
STX17	0	4	0	2	2	2	3	1	14
ABCB11	1	1	2	0	2	1	4	2	13
AR	3	2	1	0	4	0	1	2	13
BCL11A	3	1	1	0	3	1	1	3	13
CYP19A1	1	2	2	2	2	2	0	2	13
EP300	2	2	3	1	1	2	1	1	13
FKBP1A	1	1	3	1	2	2	1	2	13

	BLA T1	BLA M1	BLB T1	BLB M1	BLB M2	BH T1	BH T3	BH M2	Total
HNF1A	2	1	2	3	2	1	1	1	13
LMX1B	1	1	2	1	2	2	2	2	13
NFKB1	1	2	0	3	2	2	2	1	13
NOS3	0	1	2	2	3	1	2	2	13
NR2E1	2	3	1	1	1	3	1	1	13
NR2E3	1	1	2	3	0	2	3	1	13
RARB	3	0	3	2	1	2	1	1	13
RARG	3	2	1	3	2	1	0	1	13
RORB	2	1	1	2	1	2	2	2	13
RXRG	4	1	3	0	1	3	1	0	13
CTNNB1	1	1	3	0	1	2	2	2	12
ESRRG	1	2	1	2	2	2	0	2	12
FKBP4	2	1	2	2	1	2	1	1	12
FOXO1	0	1	1	2	0	3	4	1	12
GCH1	1	2	1	0	2	2	2	2	12
KLF4	1	1	1	2	2	2	1	2	12
MED1	1	1	1	2	1	3	1	2	12
NCOA1	3	2	3	0	1	0	2	1	12
NPAS2	0	2	0	1	3	2	2	2	12
NR1D1	2	0	2	3	1	2	2	0	12
NR3C1	1	1	0	2	2	4	0	2	12
POU5F1	2	0	3	1	1	1	4	0	12
RARA	0	3	2	0	3	0	2	2	12
RXRB	1	2	2	0	1	1	3	2	12
AKT1	3	1	2	2	1	2	0	0	11
ESR1	2	1	1	0	2	1	2	2	11
FKBP1B	1	1	1	2	1	2	1	2	11
FOXL2	1	1	1	2	2	2	0	2	11
HIVEP1	2	2	1	0	1	2	1	2	11
LMXIA		1	1	2	1	1	4	0	11
MAPKI	2	1	2	3	1	1	0	1	11
NCOA6	3		1	3		1	0		11
PCK2	2		3		2	2	0	2	11
SGKI		3	0	2	3	0	0	3	11
SMADS		1	2			0	2	1	11
SOX2	2	2	3	0		1	1	2	11
ABCG8	1	2	0	2		2	2	1	10
CCND1	1	1	1	0	1	3	1	2	10
ESRRB	1	1	1	2	1	2	2	0	10
HDAC1	1	0	1	2	0	3	1	2	10
HDAC3	1	2	1	0	2	1	2	1	10
NCOR1	0	1	0	0	1	2	2	4	10
NR0B1	4	0	1	2	0	2	1	0	10
NR1I2	0	1	1	2	1	0	5	0	10
PITX3	1	1	1	4	0	1	2	0	10
RPS27A	3	1	1	0	2	2	1	0	10
TAB2	0	3	0	3	2	0	0	2	10
TRERF1	0	1	0	1	2	3	2	1	10
ARNTL	1	1	1	0	1	1	2	2	9
BCL6	1	2	1	1	1	0	2	1	9
CDK2	1	2	1	0	2	0	0	3	9
CYP2E1	3	1	1	1	1	1	0	1	9
ESR2	3	0	1	2	0	1	1	1	9
INS	1	0	3	0	0	5	0	0	9
MEF2D	1	2	1	2	2	0	0	1	9
RELA	2	1	2	0	0	2	1	1	9
RGL2	1	1	1	1	1	2	1	1	9
VANGL2	0	3	0	0	3	0	0	3	9
VDR	1	1	2	0	2	0	2	1	9
BCL2	0	1	0	1	2	1	0	3	8

	BLA T1	BLA M1	BLB T1	BLB M1	BLB M2	BH T1	ВН ТЗ	BH M2	Total
CYP1A2	1	0	2	0	0	1	4	0	8
NR2C2AP	0	2	0	3	2	0	0	1	8
NR2F1	1	0	1	2	1	1	0	2	8
NRF1	2	1	0	0	3	0	1	1	8
PPARA	0	1	0	1	1	2	1	2	8
PPARD	1	1	0	0	3	0	2	1	8
RXRA	0	0	1	1	2	1	3	0	8
CREB1	2	0	1	1	0	0	1	2	7
DR1	1	0	1	0	1	1	2	1	7
JAZF1	1	0	1	1	0	2	1	1	7
NCOA3	0	1	2	0	1	2	1	0	7
NISCH	1	0	0	1	0	2	3	0	7
PSMD2	1	1	0	1	2	0	0	2	7
SKP1	0	2	0	0	2	0	0	3	7
SMAD4	1	1	1	0	1	1	1	1	7
CLOCK	1	1	1	0	1	1	1	0	6
KAT2B	0	1	0	4	1	0	0	0	6
NR1D2	1	2	0	1	1	0	0	1	6
NR2C2	0	2	1	1	1	0	1	0	6
NRBP1	1	1	1	1	0	1	1	0	6
NRIP1	1	0	1	0	1	1	1	1	6
TCF12	2	0	1	1	0	0	1	1	6
TFG	1	1	0	0	1	2	1	0	6
NANOG	1	0	1	0	0	0	3	0	5
NR1H2	1	0	1	2	0	1	0	0	5
SMARCA2	0	0	0	2	1	0	2	0	5
SCNN1A	0	0	0	0	0	1	3	0	4
HNF4G	1	0	0	0	0	1	1	0	3
ABCA1	1	0	0	1	0	0	0	0	2
CYP2B6	0	0	2	0	0	0	0	0	2
TCF3	1	0	0	0	0	1	0	0	2
AGO2	0	0	0		0	0	0	0	0

#### Appendix G

# Networks with Different Thresholdings

In order to identify an empirical thresholding for creating informative partial correlation networks, a series of threshold levels, ranging from 0.5% to 15% were tested. Marginalised thresholding of 1% thresholding was used in this research. Loosening the threshold (greater than 1%) increases the complexity and therefore the motif enumeration would have to be automated (eg with mfinder). On the other hand, if thresholding is tighter (less than 1%) the network would break down into fragments and not enough edges would be included to reveal an informative topology. In the following graphs, FOS and STAT1, the two hub nodes revealed from 1% thresholding partial correlation network, are coloured in yellow and the rest of the nodes are coloured in light blue. Edges are coloured with negative correlation in black and positive correlation in red. Data from TCGA class 1 from the Basal vs Luminal A analysis were used for demonstration purpose.



Figure G.1: The partial correlation network constructed using a 0.5% thresholding, containing 72 edges.



Figure G.2: The partial correlation network constructed using a 1% thresholding, containing 145 edges.



Figure G.3: The partial correlation network constructed using a 2% thresholding, containing 291 edges.



Figure G.4: The partial correlation network constructed using a 5% thresholding, containing 727 edges.



Figure G.5: The partial correlation network constructed using a 10% thresholding, containing 1453 edges.



Figure G.6: The partial correlation network constructed using a 15% thresholding, containing 2180 edges.

#### Appendix H

## Highly Expressed Genes in Basal-like Breast Cancer

Analysis	Gene
Basal-like vs Luminal A	NPAS2
	NR3C2
	PPARGC1A
	PPARGC1B
	TAF15
	FABP6
	WT1
	TMPRSS2
	STAR
	SOX9
	STAT1
	RARB
	BCL11A
	LCK
	NRP2
	POMC
	TCF3
	SIX3
	NR2E1
	NR2C2
	PPARA
	PAX6
	NR1I2

Table H.1: Genes highly expressed in Basal-like breast cancer when compared to Luminal A.

Analysis	Gene
Basal-like vs Luminal B	NR3C2
	NPAS2
	PPARA
	PAX6
	POMC
	STAR
	TAF15
	FABP6
	TMPRSS2
	PPARGC1A
	PPARGC1B
	RARB
	BCL11A
	SOX9
	WT1
	LCK
	STAT1
	STX17
	NR0B1
	NR2C2
	$\operatorname{PML}$
	NR1I2
	NR2E1
	NR4A3
	NR4A1
	CDK2
	TCF3

Table H.2: Genes highly expressed in Basal-like breast cancer when compared to Luminal B.

Analysis	Gene
Basal-like vs Her2	NR3C2
	TAF15
	FABP1
	TMPRSS2
	LPL
	SOX9
	NPAS2
	FABP6
	PPARGC1A
	STAR
	PPARA
	POMC
	PPARGC1B
	RARB
	PAX6
	WT1
	STAT1
	BCL11A
	LCK
	CYP2E1
	SMARCA2
	FOXL2
	NR4A3
	NR1I3
	NOS3
	NR4A1
	SRC
	STX17
	NRBP1
	RPS27A
	MAPK1
	RXRB
	FKBP1A
	NRP2
	SIX3
	PSMD2
	SMARCA4
	TCF3
	CDK2
	NR1I2
	NR2E1
	NB2C2
	PML
	1 1111

Table H.3: Genes highly expressed in Basal-like breast cancer when compared to Her2.

#### Appendix I

### **Docking Flags**

This appendix shows the flags and parameters used for running the Rosetta refinement docking protocol. An ensemble of 1000 models were generated for each docking run.

#io flags:-s prepackstructure.pkk.pdb -native 2iuh.pdb  $-\mathrm{out:pdb}$ -scorefile noLow\_score.sc #number of structures to produce -nstruct 1000 #flexpepdock flags -flexPepDocking:flexpep\_score\_only -flexPepDocking:pep\_refine #packing flags -ex1-ex2aro # use receptor side chains information -unboundrot 2iuh\_apo.pdb -use\_input\_sc #mute logging: -mute protocols.moves.RigidBodyMover -mute core.chemical -mute core.scoring.etable -mute protocols.evalution -mute core.pack.rotamer\_trials -mute protocols.abinitio.FragmentMover -mute core.fragment

-mute protocols.jd2.PDBJobInputter