

Contents

Glossary	xiii
Acronyms	xv
1 Introduction and Literature Review	1
1.1 Cancer Research in the Post-Genomic Era	1
1.1.1 Cancer is a Global Health Issue	2
1.1.1.1 The Genetics and Molecular Biology of Cancers	3
1.1.2 The Genomics Revolution in Cancer Research	3
1.1.2.1 High-Throughput Technologies	4
1.1.2.2 Bioinformatics and Genomic Data	5
1.1.3 Genomics Projects	5
1.1.3.1 The Cancer Genome Project	6
1.1.3.2 The Cancer Genome Atlas Project	6
1.1.4 Genomic Cancer Medicine	8
1.1.4.1 Cancer Genes and Driver Mutations	8
1.1.4.2 Precision Cancer Medicine	9
1.1.4.3 Molecular Diagnostics and Pan-Cancer Medicine	9
1.1.4.4 Targeted Therapeutics and Pharmacogenomics	10
1.1.5 Systems and Network Biology	11
1.2 Synthetic Lethal Cancer Medicine	12
1.2.1 Synthetic Lethal Genetic Interactions	13
1.2.2 Synthetic Lethal Concepts in Genetics	13
1.2.3 Synthetic Lethality in Model Systems	15
1.2.3.1 Synthetic Lethal Pathways and Networks	15
1.2.3.2 Evolution of Synthetic Lethality	16
1.2.4 Synthetic Lethality in Cancer	17
1.2.5 Clinical Impact of Synthetic Lethality in Cancer	18
1.2.6 High-throughput Screening for Synthetic Lethality	20
1.2.6.1 Synthetic Lethal Screens	21
1.2.7 Computational Prediction of Synthetic Lethality	22
1.2.7.1 Bioinformatics Approaches to Genetic Interactions	22
1.2.7.2 Comparative Genomics	24
1.2.7.3 Analysis and Modelling of Protein Data	26
1.2.7.4 Differential Gene Expression	28
1.2.7.5 Data Mining and Machine Learning	29

1.2.7.6	Mutually Exclusive Bimodality	31
1.2.7.7	Rationale for Further Development	32
1.3	E-cadherin as a Synthetic Lethal Target	33
1.3.1	The <i>CDH1</i> gene and its Biological Functions	33
1.3.1.1	Cytoskeleton	33
1.3.1.2	Extracellular and Tumour Micro-environment	34
1.3.1.3	Cell-Cell Adhesion and Signalling	34
1.3.2	<i>CDH1</i> as a Tumour (and Invasion) Suppressor	34
1.3.2.1	Breast Cancers and Invasion	35
1.3.3	Hereditary Diffuse Gastric (and Lobular Breast) Cancer	35
1.3.4	Cell Line Models of <i>CDH1</i> Null Mutations	36
1.4	Summary and Research Direction of Thesis	37
1.4.1	Thesis Aims	39
2	Methods and Resources	40
2.1	Bioinformatics Resources for Genomics Research	40
2.1.1	Public Data and Software Packages	40
2.1.1.1	Cancer Genome Atlas Data	41
2.1.1.2	Reactome and Annotation Data	42
2.2	Data Handling	42
2.2.1	Normalisation	42
2.2.2	Sample Triage	43
2.2.3	Metagenes and the Singular Value Decomposition	43
2.2.4	Candidate Triage and Integration with Screen Data	45
2.3	Techniques	46
2.3.1	Statistical Procedures and Tests	46
2.3.2	Gene Set Over-representation Analysis	47
2.3.3	Clustering	47
2.3.4	Heatmap	47
2.3.5	Modelling and Simulations	48
2.3.5.1	Receiver Operating Characteristic Curves	49
2.3.6	Resampling Analysis	49
2.4	Pathway Structure Methods	50
2.4.1	Network and Graph Analysis	50
2.4.2	Sourcing Graph Structure Data	51
2.4.3	Constructing Pathway Subgraphs	51
2.4.4	Network Analysis Metrics	52
2.5	Implementation	53
2.5.1	Computational Resources and Linux Utilities	53
2.5.2	R Language and Packages	54
2.5.3	High Performance and Parallel Computing	57
3	Methods Developed During Thesis	59
3.1	A Synthetic Lethal Detection Methodology	59
3.2	Synthetic Lethal Simulation and Modelling	61
3.2.1	A Model of Synthetic Lethality in Expression Data	62

3.2.2	Simulation Procedure	66
3.3	Detecting Simulated Synthetic Lethal Partners	68
3.3.1	Binomial Simulation of Synthetic Lethality	69
3.3.2	Multivariate Normal Simulation of Synthetic Lethality	71
3.3.2.1	Multivariate Normal Simulation with Correlated Genes	73
3.3.2.2	Specificity with Query-Correlated Pathways	80
3.4	Graph Structure Methods	83
3.4.1	Upstream and Downstream Gene Detection	83
3.4.1.1	Permutation Analysis for Statistical Significance	84
3.4.1.2	Hierarchy Based on Biological Context	84
3.4.2	Simulating Gene Expression from Graph Structures	85
3.5	Customised Functions and Packages Developed	89
3.5.1	Synthetic Lethal Interaction Prediction Tool	90
3.5.2	Data Visualisation	90
3.5.3	Extensions to the iGraph Package	92
3.5.3.1	Sampling Simulated Data from Graph Structures	92
3.5.3.2	Plotting Directed Graph Structures	93
3.5.3.3	Computing Information Centrality	94
3.5.3.4	Testing Pathway Structure with Permutation Testing	94
3.5.3.5	Metapackage to Install iGraph Functions	94
4	Synthetic Lethal Analysis of Gene Expression Data	95
4.1	Synthetic Lethal Genes in Breast Cancer	96
4.1.1	Synthetic Lethal Pathways in Breast Cancer	98
4.1.2	Expression Profiles of Synthetic Lethal Partners	99
4.1.2.1	Subgroup Pathway Analysis	102
4.2	Comparing Synthetic Lethal Gene Candidates	105
4.2.1	Primary siRNA Screen Candidates	105
4.2.2	Comparison with Correlation	106
4.2.3	Comparison with Primary Screen Viability	108
4.2.4	Comparison with Secondary siRNA Screen Validation	109
4.2.5	Comparison to Primary Screen at Pathway Level	111
4.2.5.1	Resampling Genes for Pathway Enrichment	113
4.2.6	Integrating Synthetic Lethal Pathways and Screens	116
4.3	Metagene Analysis	118
4.3.1	Pathway Expression	119
4.3.2	Somatic Mutation	121
4.3.3	Synthetic Lethal Pathway Metagenes	125
4.3.4	Synthetic Lethality in Breast Cancer	126
4.4	Replication in Stomach Cancer	127
4.5	Discussion	128
4.5.1	Strengths of the SLIPT Methodology	128
4.5.2	Synthetic Lethal Pathways for E-cadherin	129
4.5.3	Replication and Validation	131
4.5.3.1	Integration with short interfering RNA (siRNA) Screen- ing	131

4.5.3.2	Replication across Tissues	132
4.6	Summary	132
5	Synthetic Lethal Pathway Structure	134
5.1	Synthetic Lethal Genes in Reactome Pathways	134
5.1.1	The PI3K/AKT Pathway	135
5.1.2	The Extracellular Matrix	137
5.1.3	G Protein Coupled Receptors	140
5.1.4	Gene Regulation and Translation	140
5.2	Network Analysis of Synthetic Lethal Genes	141
5.2.1	Gene Connectivity and Vertex Degree	142
5.2.2	Gene Importance and Centrality	143
5.2.2.1	Information Centrality	143
5.2.2.2	PageRank Centrality	145
5.3	Relationships between Synthetic Lethal Genes	147
5.3.1	Hierarchical Pathway Structure	147
5.3.1.1	Contextual Hierarchy of PI3K	147
5.3.1.2	Testing Contextual Hierarchy of Synthetic Lethal Genes	147
5.3.2	Upstream or Downstream Synthetic Lethality	151
5.3.2.1	Measuring Structure of Candidates within PI3K	151
5.3.2.2	Resampling for Synthetic Lethal Pathway Structure . .	153
5.4	Discussion	155
5.5	Summary	157
6	Simulation and Modelling of Synthetic Lethal Pathways	159
6.1	Synthetic Lethal Detection Methods	160
6.1.1	Performance of SLIPT and χ^2 across Quantiles	160
6.1.1.1	Correlated Query Genes affects Specificity	164
6.1.2	Alternative Synthetic Lethal Detection Strategies	166
6.1.2.1	Correlation for Synthetic Lethal Detection	166
6.1.2.2	Testing for Bimodality with BiSEp	168
6.2	Simulations with Graph Structures	169
6.2.1	Performance over Graph Structures	170
6.2.1.1	Simple Graph Structures	170
6.2.1.2	Constructed Graph Structures	172
6.2.2	Performance with Inhibitions	175
6.2.3	Synthetic Lethality across Graph Structures	180
6.2.4	Performance within a Simulated Human Genome	184
6.3	Simulations in More Complex Graph Structures	188
6.3.1	Simulations over Pathway-based Graphs	189
6.3.2	Pathway Structures in a Simulated Human Genome	192
6.4	Discussion	195
6.4.1	Simulation Procedure	195
6.4.2	Comparing Methods with Simulated Data	196
6.4.3	Design and Performance of SLIPT	197
6.4.4	Simulations from Graph Structures	199

6.5	Summary	200
7	Discussion	201
7.1	Synthetic Lethality and <i>CDH1</i> Biology	201
7.1.1	Established Functions of <i>CDH1</i>	202
7.1.2	The Molecular Role of <i>CDH1</i> in Cancer	202
7.2	Significance	203
7.2.1	Synthetic Lethality in the Genomic Era	203
7.2.2	Clinical Interventions based on Synthetic Lethality	205
7.3	Future Directions	206
7.4	Conclusions	208
	Bibliography	210
A	Sample Quality	234
A.1	Sample Correlation	234
A.2	Replicate Samples in The Cancer Genome Atlas (TCGA) Breast	236
B	Software Used for Thesis	240
C	Mutation Analysis in Breast Cancer	249
C.1	Synthetic Lethal Genes and Pathways	249
C.2	Synthetic Lethal Expression Profiles	250
C.3	Comparison to Primary Screen	253
C.3.1	Resampling Analysis	255
C.4	Compare Synthetic Lethal Interaction Prediction Tool (SLIPT) genes	257
C.5	Metagene Analysis	259
C.6	Expression of Somatic Mutations	260
C.7	Metagene Expression Profiles	263
D	Intrinsic Subtyping	266
E	Stomach Expression Analysis	268
E.1	Synthetic Lethal Genes and Pathways	268
E.2	Comparison to Primary Screen	272
E.2.1	Resampling Analysis	274
E.3	Metagene Analysis	276
F	Synthetic Lethal Genes in Pathways	277
G	Pathway Connectivity for Mutation SLIPT	285
H	Information Centrality for Gene Essentiality	289
I	Pathway Structure for Mutation SLIPT	292
J	Performance of SLIPT and χ^2	295
J.1	Correlated Query Genes affects Specificity	301

K	Simulations on Graph Structures	307
K.0.1	Simulations from Inhibiting Graph Structures	308
K.1	Simulation across Graph Structures	311
K.2	Simulations from Complex Graph Structures	315
K.2.1	Simulations from Complex Inhibiting Graphs	318
K.3	Simulations from Pathway Graph Structures	324

List of Figures

1.1	Synthetic genetic interactions	14
1.2	Synthetic lethality in cancer	17
2.1	Read count density	44
2.2	Read count sample mean	44
3.1	Framework for synthetic lethal prediction	60
3.2	Synthetic lethal prediction adapted for mutation	61
3.3	A model of synthetic lethal gene expression	63
3.4	Modelling synthetic lethal gene expression	64
3.5	Synthetic lethality with multiple genes	65
3.6	Simulating gene function	67
3.7	Simulating synthetic lethal gene function	67
3.8	Simulating synthetic lethal gene expression	68
3.9	Performance of binomial simulations	70
3.10	Comparison of statistical performance	70
3.11	Performance of multivariate normal simulations	72
3.12	Simulating expression with correlated gene blocks	74
3.13	Simulating expression with correlated gene blocks	75
3.14	Synthetic lethal prediction across simulations	76
3.15	Performance with correlations	77
3.16	Comparison of statistical performance with correlation structure	78
3.17	Performance with query correlations	79
3.18	Statistical evaluation of directional criteria	81
3.19	Performance of directional criteria	82
3.20	Simulated graph structures	86
3.21	Simulating expression from a graph structure	87
3.22	Simulating expression from graph structure with inhibitions	88
3.23	Demonstration of violin plots with custom features	91
3.24	Demonstration of annotated heatmap	91
3.25	Simulating graph structures	93
4.1	Synthetic lethal expression profiles of analysed samples	101
4.2	Comparison of SLIPT to siRNA	105
4.3	Compare SLIPT and siRNA genes with correlation	106
4.4	Compare SLIPT and siRNA genes with correlation	107
4.5	Compare SLIPT and siRNA genes with viability	108

4.6	Compare SLIPT genes with siRNA viability	109
4.7	Resampled intersection of SLIPT and siRNA candidates	113
4.8	Pathway metagene expression profiles	120
4.9	Expression profiles for constituent genes of PI3K	122
4.10	Expression profiles for estrogen receptor related genes	123
4.11	Somatic mutation against the PI3K metagene	124
5.1	synthetic lethality in the PI3K cascade	136
5.2	synthetic lethality in Elastic Fibre Formation	138
5.3	Synthetic lethality in Fibrin Clot Formation	139
5.4	Synthetic lethality and vertex degree	142
5.5	Synthetic lethality and centrality	145
5.6	Synthetic lethality and PageRank	146
5.7	Hierarchical structure of PI3K	148
5.8	Hierarchy score in PI3K against synthetic lethality in PI3K	149
5.9	Structure of synthetic lethality in PI3K	151
5.10	Structure of synthetic lethality resampling in PI3K	152
6.1	Performance of χ^2 and SLIPT across quantiles	162
6.2	Performance of χ^2 and SLIPT across quantiles with more genes	163
6.3	Performance of χ^2 and SLIPT across quantiles with query correlation	164
6.4	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	165
6.5	Performance of negative correlation and SLIPT	167
6.6	Simple graph structures	170
6.7	Performance of simulations on a simple graph	171
6.8	Performance of simulations is similar in simple graphs	173
6.9	Performance of simulations on a pathway	174
6.10	Performance of simulations on a simple graph with inhibition	176
6.11	Performance is higher on a simple inhibiting graph	177
6.12	Performance of simulations on a constructed graph with inhibition	178
6.13	Performance is affected by inhibition in graphs	180
6.14	Detection of synthetic lethality within a graph structure	182
6.15	Performance of simulations including a simple graph	185
6.16	Performance on a simple graph improves with more genes	186
6.17	Performance on an inhibiting graph improves with more genes	188
6.18	Performance of simulations on the PI3K cascade	191
6.19	Performance of simulations including the PI3K cascade	193
6.20	Performance on pathways improves with more genes	194
A.1	Correlation profiles of removed samples	234
A.2	Correlation analysis and sample removal	235
A.3	Replicate excluded samples	236
A.4	Replicate samples with all remaining	237
A.5	Replicate samples with some excluded	238
C.1	Synthetic lethal expression profiles of analysed samples	251

C.2	Comparison of mtSLIPT to siRNA	253
C.3	Compare mtSLIPT and siRNA genes with correlation	257
C.4	Compare mtSLIPT and siRNA genes with correlation	257
C.5	Compare mtSLIPT and siRNA genes with siRNA viability	258
C.6	Somatic mutation against PIK3CA metagene	260
C.7	Somatic mutation against PI3K protein	261
C.8	Somatic mutation against AKT protein	262
C.9	Pathway metagene expression profiles	263
C.10	Expression profiles for p53 related genes	264
C.11	Expression profiles for BRCA related genes	265
E.1	Synthetic lethal expression profiles of stomach samples	270
E.2	Comparison of SLIPT in stomach to siRNA	272
F.1	Synthetic lethality in the PI3K/AKT pathway	277
F.2	Synthetic lethality in the PI3K/AKT pathway in cancer	278
F.3	Synthetic lethality in the Extracellular Matrix	279
F.4	Synthetic lethality in the GPCRs	280
F.5	Synthetic lethality in the GPCR Downstream	281
F.6	Synthetic lethality in the Translation Elongation	282
F.7	Synthetic lethality in the Nonsense-mediated Decay	283
F.8	Synthetic lethality in the 3' UTR	284
G.1	Synthetic lethality and vertex degree	285
G.2	Synthetic lethality and centrality	286
G.3	Synthetic lethality and PageRank	287
H.1	Information centrality distribution	291
I.1	Synthetic lethality and heirarchy score in PI3K	292
I.2	Heirarchy score in PI3K against synthetic lethality in PI3K	293
I.3	Structure of synthetic lethality in PI3K	293
I.4	Structure of synthetic lethality resampling	294
J.1	Performance of χ^2 and SLIPT across quantiles	295
J.2	Performance of χ^2 and SLIPT across quantiles	297
J.3	Performance of χ^2 and SLIPT across quantiles with more genes	299
J.4	Performance of χ^2 and SLIPT across quantiles with query correlation	301
J.5	Performance of χ^2 and SLIPT across quantiles with query correlation	303
J.6	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	305
K.1	Performance of simulations on a simple graph	307
K.2	Performance of simulations on an inhibiting graph	308
K.3	Performance of simulations on a constructed graph with inhibition	309
K.4	Performance of simulations on a constructed graph with inhibition	310
K.5	Detection of synthetic lethality within a graph structure	311
K.6	Detection of synthetic lethality within an inhibiting graph	313

K.7	Detection of synthetic lethality within an inhibiting graph	314
K.8	Performance of simulations on a branching graph	315
K.9	Performance of simulations on a complex graph	316
K.10	Performance of simulations on a large graph	317
K.11	Performance of simulations on a branching graph with inhibition	318
K.12	Performance of simulations on a branching graph with inhibition	319
K.13	Performance of simulations on a complex graph with inhibition	320
K.14	Performance of simulations on a complex graph with inhibition	321
K.15	Performance of simulations on a large constructed graph with inhibition	322
K.16	Performance of simulations on a large constructed graph with inhibition	323
K.17	Performance of simulations on the $G_{\alpha i}$ signalling pathway	324
K.18	Performance of simulations including the $G_{\alpha i}$ signalling pathway	325

List of Tables

1.1	Methods for predicting genetic interactions	23
1.2	Methods for predicting synthetic lethality in cancer	23
1.3	Methods used by Wu <i>et al.</i> (2014)	25
2.1	Excluded samples by batch and clinical characteristics.	43
2.2	Computers used during thesis	53
2.3	Linux utilities and applications used during thesis	54
2.4	R installations used during thesis	55
2.5	R Packages used during thesis	55
2.6	R packages developed during thesis	57
4.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT	97
4.2	Pathways for <i>CDH1</i> partners from SLIPT	99
4.3	Pathways for clusters of <i>CDH1</i> partners from SLIPT	103
4.4	ANOVA for synthetic lethality and correlation with <i>CDH1</i>	107
4.5	Comparing SLIPT genes against secondary siRNA screen	110
4.6	Pathways for <i>CDH1</i> partners from SLIPT and siRNA	112
4.7	Pathways for <i>CDH1</i> partners from SLIPT	115
4.8	Pathways for <i>CDH1</i> partners from SLIPT and siRNA primary screen .	117
4.9	Candidate synthetic lethal metagenes against <i>CDH1</i> from SLIPT . . .	126
5.1	ANOVA for synthetic lethality and vertex degree	143
5.2	ANOVA for synthetic lethality and information centrality	145
5.3	ANOVA for synthetic lethality and PageRank centrality	147
5.4	ANOVA for synthetic lethality and PI3K hierarchy	150
5.5	Resampling for pathway structure of synthetic lethal detection methods	154
B.1	Complete list of R packages used during this thesis	240
C.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from mtSLIPT . . .	249
C.2	Pathways for <i>CDH1</i> partners from mtSLIPT	250
C.3	Pathways for clusters of <i>CDH1</i> partners from mtSLIPT	252
C.4	Pathways for <i>CDH1</i> partners from mtSLIPT and siRNA	254
C.5	Pathways for <i>CDH1</i> partners from mtSLIPT	255
C.6	Pathways for <i>CDH1</i> partners from mtSLIPT and siRNA primary screen	256
C.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from mtSLIPT . .	259
D.1	Comparison of intrinsic subtypes	266

E.1	Synthetic lethal gene partners of <i>CDH1</i> from SLIPT in stomach cancer	268
E.2	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	269
E.3	Pathways for clusters of <i>CDH1</i> partners in stomach SLIPT	271
E.4	Pathways for <i>CDH1</i> partners from SLIPT and siRNA	273
E.5	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	274
E.6	Pathways for <i>CDH1</i> partners from SLIPT in stomach and siRNA . . .	275
E.7	Synthetic lethal metagenes against <i>CDH1</i> in stomach cancer	276
G.1	ANOVA for synthetic lethality and vertex degree	288
G.2	ANOVA for synthetic lethality and information centrality	288
G.3	ANOVA for synthetic lethality and PageRank centrality	288
H.1	Information centrality for genes and molecules in the Reactome network	290
I.1	ANOVA for synthetic lethality and PI3K hierarchy	292
I.2	Resampling for pathway structure of synthetic lethal detection methods	294

Glossary

bioinformatics	Statistical or computational approaches to biological data or research tools.
E-cadherin	Epithelial cadherin (calcium-dependent adhesion), a cell-adhesion protein encoded by <i>CDH1</i> .
gene expression	A measure of the relative expression of each gene from the mRNA extracted from (pooled) cells.
genome	All of the DNA sequence in the genome.
genomic	The use of data from all genes in the genome.
graph or network	A mathematical structure modelling or depicting the relationships between elements.
metagene	A consistent signal of expression for a collection of genes such as a biological pathway, derived from singular value decomposition.
microarray	A high-throughput technique to measure presence or abundance of nucleic acid sequences from binding to probes.
mutation	A change in DNA sequence that disrupts gene function.
RNA-Seq	The generation of transcriptome data from sequencing RNA.
synthetic lethal	Genetic interactions where inactivation of multiple genes is inviable (or deleterious) which are viable if inactivated separately.
tumour suppressor	A gene potentially causes cancer, typically by disruption of functions which protect the cell from cancer.

Acronyms

ANOVA	Analysis of Variance.
AUROC	Area Under the Receiver Operating Characteristic (curve).
BiSEp	Bimodal Subsetting Expression.
DNA	Deoxyribonucleic Acid.
FDR	False Discovery Rate.
HPC	High Performance Computing.
mtSLIPT	Synthetic Lethal Interaction Prediction Tool (against mutation).
NeSI	New Zealand eScience Infrastructure.
PI3K	Phosphoinositide 3-kinase.
ROC	Receiver Operating Characteristic (curve).
siRNA	Short Interfering RNA.
SLIPT	Synthetic Lethal Interaction Prediction Tool.
Slurm	Simple Linux Utility for Resource Management.
TCGA	The Cancer Genome Atlas (genomics project).

Chapter 6

Simulation and Modelling of Synthetic Lethal Pathways

Simulation and modelling of [synthetic lethality](#) in [gene expression](#) was revisited in greater detail in this chapter, building upon the results which supported the use of [SLIPT](#) (in Section 3.3). In Chapter 3, a simulation procedure for generating simulated data with underlying (known) [synthetic lethal](#) partners of a query gene, such as *CDH1*, was developed (as described in Section 3.2.2) by sampling from a Multivariate normal distribution based on a statistical model of [synthetic lethality](#) in [expression](#) data (as described in Section 3.2.1). This simulation framework was applied to simulated data (in Section 3.3), including simple correlation structures to assess the statistical performance of the [SLIPT](#) methodology and support its use as a computational approach for detecting [synthetic lethal](#) candidates from [expression](#) data throughout this thesis (Chapters 4 and 5).

While this basic framework was provided some support for the use of [SLIPT](#), further investigations with simulations were conducted to assess the strengths and limitations of the [SLIPT](#) methodology, compare it to alternative statistical approaches to [synthetic lethal](#) detection, and assess its performance under more complex correlation structures. Together these simulation investigations assess the performance of the [SLIPT](#) methodology, including on pathway [graph](#) structures (e.g., those discussed in Chapter 5) and indicate whether the [SLIPT](#) methodology (or similar refined [bioinformatics](#) strategies) are statistically rigorous or suitable for wider [genomics](#) applications.

These simulation investigations continue to utilise the Multivariate Normal simulation procedure (as applied in Section 3.3) with further refinements. The [SLIPT](#) methodology (and the χ^2 test) were applied across a range of parameters (including

altering the quantiles for detecting [synthetic lethal](#) direction and compared to correlation). This was also applied to query correlated genes (as performed in Section 3.3).

A refined simulation procedure was developed specifically to extend the methodology described in Section 3.2 to utilise pathway [graph](#) structures for the correlation structures of simulated datasets (as described in Section 3.4.2). This methodology can be applied to simulated correlation structures across simple [graph](#) structures to test specific network modules or use [pathway](#) structures based on biological pathways. Thus [graph](#) structure and simulation approaches were combined to test whether a gene locus in a pathway affects detection by [SLIPT](#) and whether [SLIPT](#) performance is affected by [pathway](#) structure. The simulation procedure based on [graph](#) structures was applied in a computational pipeline across many parameter combinations using high-performance computing resources (as discussed in Section 2.5.3) and the core simulation functions have been released as a software package for wider use to test [bioinformatics](#) and statistical methods on [graph](#) structures (as described in Section 3.5.3).

6.1 Synthetic Lethal Detection Methods

The [SLIPT](#) methodology (as it has been applied throughout Chapters 4 and 5) was compared for alternative computational approaches to detecting [synthetic lethality](#) in simulated [gene expression](#) data. As discussed in Section 3.3, this procedure enabled testing ability of [SLIPT](#) to detect known [synthetic lethal](#) partner genes by sampling from a statistical model of [synthetic lethality](#). While comprehensive benchmarking has not been performed, several approaches to [synthetic lethal](#) detection are considered (e.g., Pearson correlation, the χ^2 test, and testing for bimodality) to evaluate the strengths of the [SLIPT](#) methodology, including modifications to the parameters of [SLIPT](#).

The following comparisons of simulations of computational detection of [synthetic lethality](#) with different statistical rationales suffice to discuss the strengths of [SLIPT](#), evaluate whether it is appropriate for further application in [genomics](#) research, and identify limitations which may be addressed with further developments. Some potential avenues for further development of computational [synthetic lethal](#) discovery will be discussed in Section 7.3.

6.1.1 Performance of [SLIPT](#) and χ^2 across Quantiles

Simulated datasets with [synthetic lethal](#) partner genes were generated using the multivariate normal simulation procedure (as described in Section 3.2.2) with performance

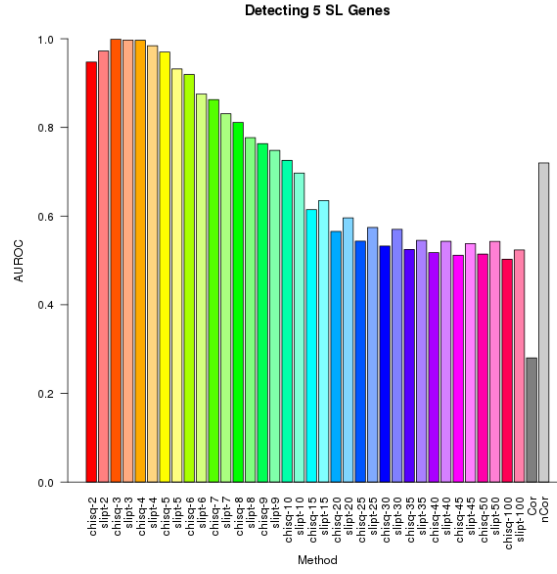
assessed using [area under receiver operating characteristic \(AUROC\)](#) analysis (as described in Section 2.3.5). [Synthetic lethal](#) detection was compared for modifications to the [SLIPT](#) methodology (as described in Section 3.1), namely that the quantiles used to define low and high [expression](#) was varied. Rather, than $1/3$ (as used throughout this thesis) the samples below the lowest $1/n$ quantile and above the highest $1/n$ quantile were used for [SLIPT](#) (and the χ^2 -test) to detect lowly and highly expressing samples respectively. The quantiles tested range from 2, splitting at the $1/2$ quantile (the median), to 100, using the lowest (1%) and highest (99%) percentiles.

This enabled testing of the threshold for lowly expressing genes which is most able to distinguish [synthetic lethal](#) genes, even with higher-order [synthetic lethal](#) interactions (as discussed in Section 3.2.1). Both [SLIPT](#) with the directional criteria for [synthetic lethality](#) and significance of the equivalent χ^2 test were performed for each quantile. Pearson correlation was also tested on simulated continuous [expression](#) data for [synthetic lethal](#) detection in simulated data, considering both positive and negative correlations separately as predictors of [synthetic lethality](#) for comparison with χ^2 based approaches, using discrete categories for gene function deriving from quantiles.

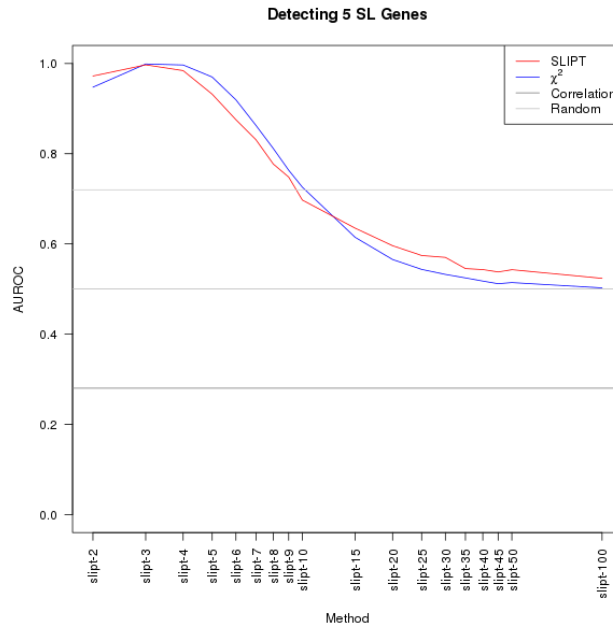
The results presented throughout this section use the example of 5 [synthetic lethal](#) partners to illustrate the differences in performance between the standard [SLIPT](#) procedure (slipt-3) to n quantiles (slipt- n), the χ^2 -test on the same quantiles, and positive or negative correlation. However, similar results across different numbers of known [synthetic lethal](#) genes are shown in Appendix J. The [synthetic lethal](#) detection procedures were compared with 10,000 simulations of a small dataset of 100 genes and 1000 samples without correlation structure between genes, as performed in Section 3.3.2). As shown in Figure 6.1, the 3-quantiles previously used have optimal performance and [SLIPT](#) has a comparable or higher performance than the χ^2 -test alone across quantiles.

Pearson correlation was also tested as a predictor of [synthetic lethality](#) (i.e., whether highly positive or negative correlations with the query gene detected [synthetic lethal](#) partners). Positive correlation performed worse than random (with an [AUROC](#) lower than 0.5) as thus coexpression of genes is not predictive of [synthetic lethality](#) in simulated data. Conversely, negative correlation is predictive of [synthetic lethality](#), consistent with [synthetic lethal](#) gene activity being mutually exclusive. However, neither correlation approach performed as well as the optimal quantiles for the [SLIPT](#) procedure or χ^2 -test.

These results are shown in both a bargraph and lineplot to show the individual results of each parameter, and to compare [SLIPT](#) with the χ^2 -test side-by-side across



(a) Barplot of χ^2 , SLIPT, and correlation.



(b) Lineplot of χ^2 , SLIPT, and correlation.

Figure 6.1: **Performance of χ^2 and SLIPT across quantiles.** Synthetic lethal detection (of 5 genes) with quantiles as on the axes. The barplot uses the same hues for each quantile (grey for correlation) and darker for χ^2 (and positive correlation). The line plot (with log-scale quantiles) is coloured according to the legend. SLIPT and χ^2 perform similarly, peaking at $1/3$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 .

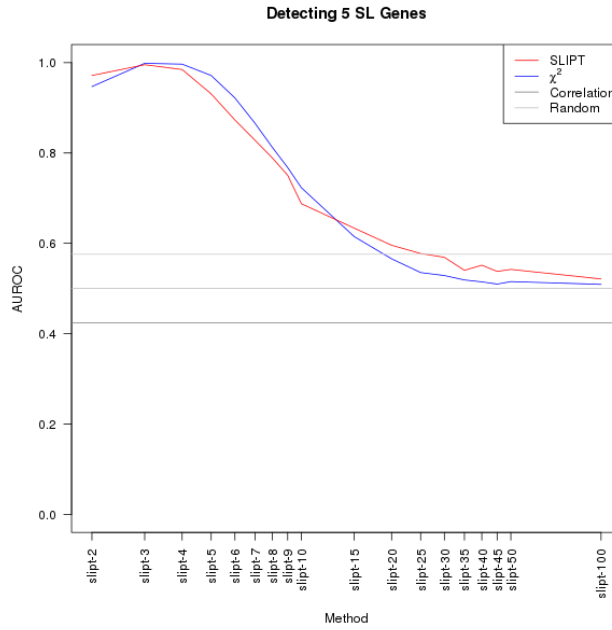


Figure 6.2: **Performance of χ^2 and SLIPT across quantiles with more genes.** **Synthetic lethal** detection (of 5 genes in 20,000) with quantiles as in axis labels. The line plot (with log-scale quantiles) is coloured according to the legend. As for simulations with fewer genes, **SLIPT** and χ^2 perform similarly, peaking at $1/3$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for **SLIPT** or χ^2 .

quantiles. Similarly, these plots are given for detecting a range of known **synthetic lethal** partners in the simulations in Figures J.1 and J.2. These demonstrate that the findings shown for 5 **synthetic lethal** genes are robust across different numbers of underlying **synthetic lethal** genes.

The **synthetic lethal** detection procedures were also tested with 1000 simulations of a larger dataset of 20,000 genes and 1000 samples. While fewer simulations gives a less accurate **receiver operating characteristic (ROC)** result, this is sufficient to replicate the above findings with a feasible number of genes in a human **gene expression** dataset and assess the impact of a higher proportion of non synthetic lethal genes (potential false positives). Simulated datasets of this size were also used in Section 3.3.2 to test the specificity in a number of genes similar to that in experimental datasets for cancer **genomes**. As shown in Figure 6.2, the above findings were replicated in simulations of a larger dataset with 20,000 genes. These were also robustly replicated across varying numbers of underlying **synthetic lethal** genes (as shown in Figure J.3).

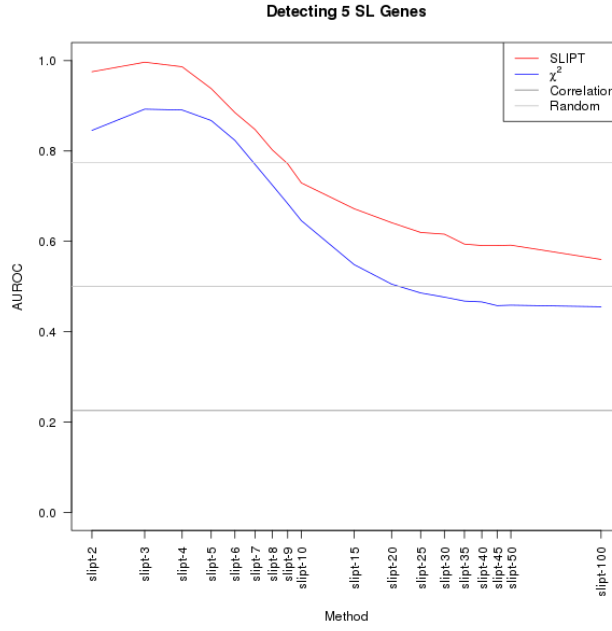


Figure 6.3: **Performance of χ^2 and SLIPT across quantiles with query correlation.** Synthetic lethal detection (of 5 genes in 100 including 5 query correlated) with quantiles as in axis labels. The line plot (with log-scale quantiles) is coloured according to the legend. SLIPT performs consistently higher than χ^2 due to higher specificity. Negative correlation performed modestly.

6.1.1.1 Correlated Query Genes affects Specificity

As discussed in Section 3.3.2.2, positively correlated genes (with the query gene) have an impact of on the performance of synthetic lethal detection. SLIPT was able to distinguish these from synthetic lethal partners and hence is likely to have a higher specificity in datasets which include positively correlated genes with the query gene (as expected in gene expression data). The synthetic lethal detection procedures were compared with 10,000 simulations of a small dataset of 100 genes (with 5 correlated with the query gene) and 1000 samples otherwise without correlation structure between genes. As shown in Figure 6.3, this specificity is reflected in the increased AUROC performance values for SLIPT (in contrast to Figure 6.1). This specificity can be attributed to the directional criteria (as described in Section 3.1) since the χ^2 -test alone performs comparatively poorly with positively correlated genes.

The synthetic lethal detection procedures were also compared with 1000 simulations of a larger dataset of 20,000 genes (with 1000 correlated with the query gene) and 1000 samples otherwise without correlation structure between genes. This simulation

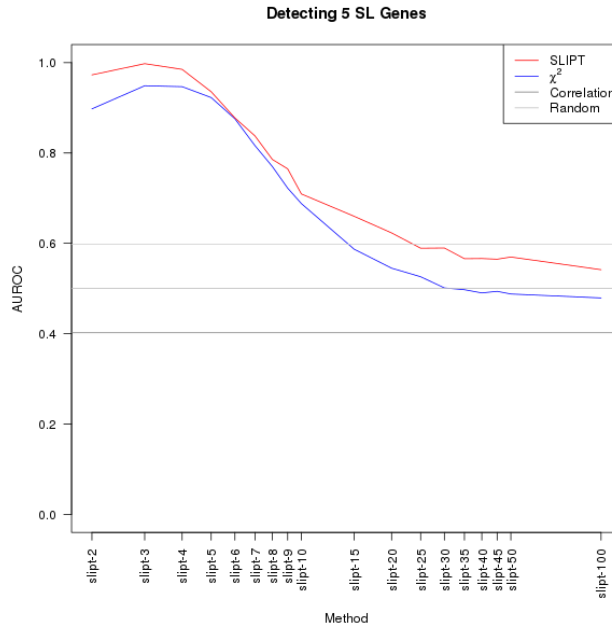


Figure 6.4: **Performance of χ^2 and SLIPT across quantiles with query correlation and more genes.** Synthetic lethal detection (of 5 genes in 20,000 including 1000 query correlated) with quantiles as in axis labels. The line plot (with log-scale quantiles) is coloured according to the legend. SLIPT performs consistently higher than χ^2 due to higher specificity. Negative correlation performed modestly.

increases the number of genes (and proportion of negative genes) to those comparable with a human gene expression dataset while maintaining a comparable 5% of positively correlated genes. As shown in Figure 6.4, SLIPT still outperforms χ^2 or negative correlation and is optimal at the 3-quantile. The difference between SLIPT and χ^2 was less pronounced in a larger dataset with many weakly correlated genes. The greater specificity of SLIPT than χ^2 -test to distinguish positively correlated non synthetic lethal genes is not as evident with a large number of negative genes (as potential false positives). However, specificity is an important consideration in large-scale genomics analysis where there are potentially many false positives.

Nevertheless, SLIPT with 3-quantiles (as performed throughout Chapters 4 and 5), had higher performance than when other quantile thresholds were used, particularly when positive correlations were present (replicating the Section 3.3.2.2). These findings hold across different numbers of underlying synthetic lethal genes (as shown in Figures J.5 and J.6).

Together these results support the use of **SLIPT**, particularly the use of quantiles as thresholds for gene function and specific use of 3-quantiles which perform well compared to other quantiles. A particular concern in the design of **SLIPT** for **expression** data whether the samples sizes are sufficient when the data is divided into quantiles. The **SLIPT** methodology further performed better for 3-quantiles (and other moderate values) than χ^2 or correlation as a predictor of **synthetic lethality**. These results are irrespective of sample size or p-value threshold since the results replicated across sample sizes and the **AUROC** values were independent significance thresholds. Using a moderate number of quantiles for **SLIPT** ensures that there are a sufficient number of samples expected below and above them so that deviations from these are statistically detectable. These quantiles were also optimal for the χ^2 test which uses the same expected values as the **SLIPT** directional conditions.

6.1.2 Alternative Synthetic Lethal Detection Strategies

The **SLIPT** approach (and χ^2) to detect **synthetic lethality** from binning **expression** to estimate gene function also outperforms correlations which use continuous data directly. Correlation performing poorly as a **synthetic lethal** detection strategy consistent with there not necessarily being a relationship between **synthetic lethal** partners which can be in distinct biological pathways, expressed at different times or in different cell types. Nevertheless, correlation is among the alternative detection methods considered in further detail.

The **BImodal Subsetting ExPression** (**BiSep**) R package (Wappett, 2014) for using bimodality to detect **synthetic lethality** (Wappett *et al.*, 2016) were also considered, along with a linear regression approach. These statistical methods span a range of computational approaches to detecting **synthetic lethality** and serve to compare alternatives to **SLIPT**, supporting its design and application. However, these comparisons are able to provide supporting data from statistical modelling and simulations for the viability of the **SLIPT** methodology for **synthetic lethal** discovery in cancer (as demonstrated in Chapter 4) and further applications.

6.1.2.1 Correlation for Synthetic Lethal Detection

As shown in Section 6.1.1, negative (Pearson) correlation performed better than positive correlation, indicating the inverse relationships were more predictive of **synthetic lethality**. However, neither correlation approach performed as well as **SLIPT** or the χ^2 test as a predictor of **synthetic lethal** gene partners. It is notable that negative correlation still often performed considerably better than random chance.

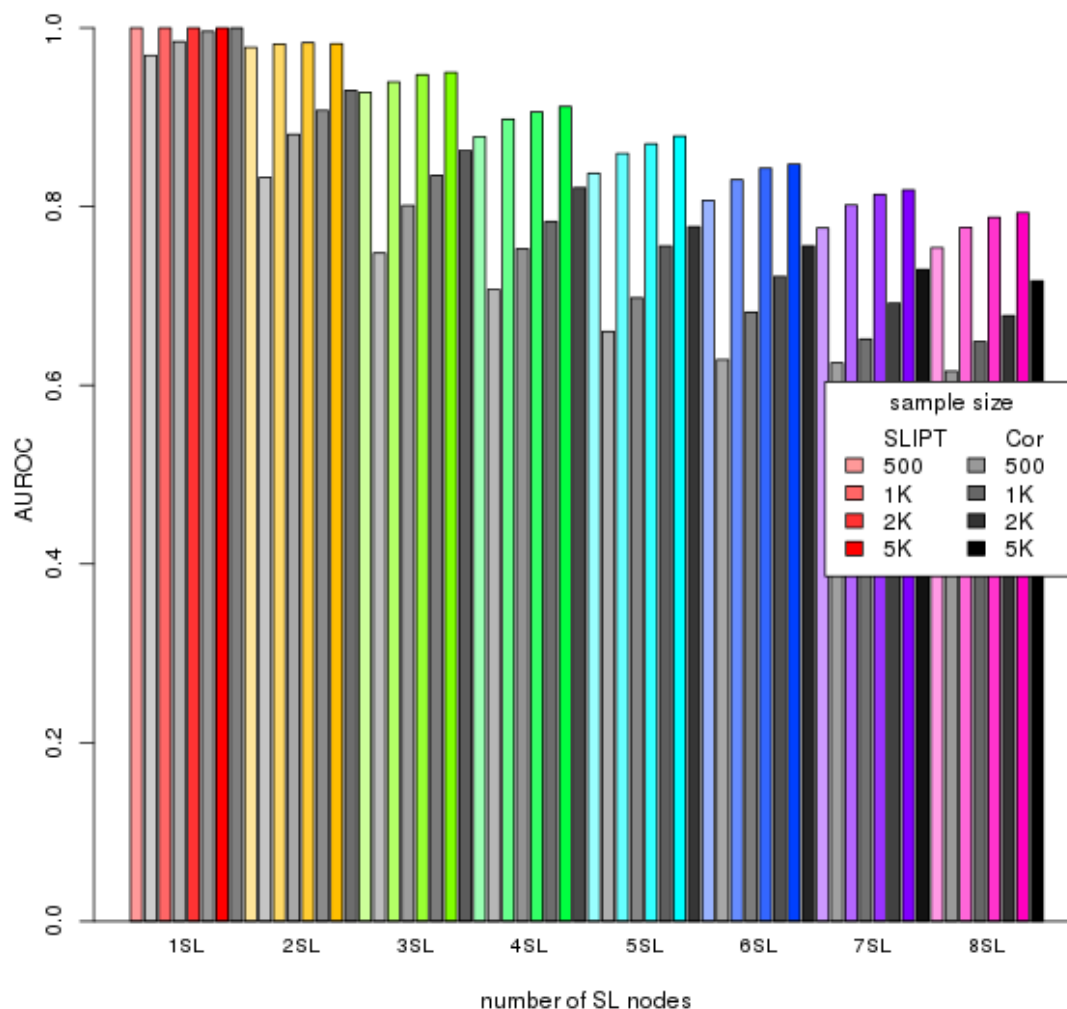


Figure 6.5: **Performance of negative correlation and SLIPT.** Synthetic lethal detection with SLIPT was compared to negative (Pearson) correlation across parameters. SLIPT consistently outperformed correlation. Both approaches had lower performance for more synthetic lethal partners and lower sample sizes. 10,000 simulations were performed with correlation structure as used in Section 6.2.1.2

Negative correlation was compared directly to the SLIPT methodology (as described in Section 3.1) across numbers of known synthetic lethal partners and sample size (ranging from 500 to 5000). This comparison used 1000 simulations of a dataset with 20,000 genes and synthetic lethal genes from within a network (sampled as in Section 3.4.2)) with a 0.8 correlation between adjacent genes. In a direct comparison of

SLIPT and negative correlation (shown in Figure 6.5), SLIPT consistently has higher performance in simulated data across parameter values and (inverse) correlation-based approaches perform modestly in comparison. Thus using thresholds to categorise expression data (as performed by SLIPT and χ^2) does not compromise the performance of these methods by losing continuous data that would be used for calculating correlations. Similarly, the slope of a linear regression did not perform as well at synthetic lethal detection than SLIPT.

Both SLIPT and correlation had poorer performance with increasing numbers of the synthetic lethal genes to detect, while they had higher performance in higher sample sizes, as expected (as previously observed for SLIPT in Section 3.3). Thus the issue with detection of greater numbers of synthetic lethal genes is not specific to SLIPT but occurs across computational methods of synthetic lethal discovery in (simulated) expression data and likely stems from cryptic higher-order synthetic lethal interactions (as conservatively assumed in Section 3.2.1).

6.1.2.2 Testing for Bimodality with BiSEp

Extensive attempts were also made to compare SLIPT to the BiSEp methodology (Wappett *et al.*, 2016), a statistical approach to identify synthetic lethal gene pairs from mutually exclusive relationships using bimodal distributions. This synthetic lethal detection methodology is also designed for expression analysis in cancer and is readily available as an (open-source) R package (Wappett, 2014), a practice which facilitates adoption and testing of the methodology on the same datasets and simulations procedures as previously used for SLIPT.

The BiSEp package is designed for global testing of all potential gene pairs in the genomes for synthetic lethality rather than focusing on the search space of potential partners of the query gene. This approach was unable to detect synthetic lethal gene pairs in the TCGA breast cancer expression dataset (TCGA, 2012). However, this may be due to stringent thresholds under the multiple testing of millions of potential gene pairs.

For a direct comparison with the query-based SLIPT approach, the source code of the BiSEp R functions was modified to test solely for the partners of a specific gene. This approach was still unable to detect synthetic lethal partners of *CDH1* in TCGA breast cancer expression data (TCGA, 2012), even with the detection thresholds for bimodality and significance greatly relaxed from those which the package defaults to.

To circumvent multiple testing issues, BiSEp only tests gene pairs for synthetic lethality between genes with a detectable bimodal distribution. However, even with

relaxed thresholds, bimodal distributions were not detectable in the normalised TCGA data (TCGA, 2012). Such normalisation Ritchie *et al.* (2015) is standard practice for expression datasets generated from microarrays or RNA-Seq and therefore BiSEp may not be appropriate to apply to this data. However, it is noted that BiSEp may also use other data types such as DNA copy number or cell line data for which it may be more applicable (Wappett *et al.*, 2016).

Nevertheless, attempts were made to test BiSEp on simulated datasets with underlying synthetic lethal genes (using the procedures described in Sections 3.2.2 and 3.4.2). However, BiSEp was also unable to detect genes with bimodal distributions of genes (and thus unable to detect synthetic lethality) in a limited number of computationally intensive simulations. Therefore investigations on a wider range of parameters were not performed.

6.2 Simulations with Graph Structures

The simulations of synthetic lethality performed in Section 3.3 included correlated blocks of genes as a rudimentary representation of pathway structure and co-regulated genes. The simulation procedure has been enhanced to account for more complex pathway structures by sampling from multivariate normal distributions with correlation structure derived from graph structures (as described in Section 3.4.2). This approach enabled the simulation of synthetic lethal pathways with known correlation structure and synthetic lethal partners (of a gene not in the pathway). Using this procedure, the performance of SLIPT was evaluated under simple controlled correlation structures and complex correlations, such as those derived from biological networks (e.g., those described in Chapter 5). The SLIPT methodology was tested in artificial constructed networks to evaluate the effect of pathway structure on synthetic lethal detection. These included large biologically feasible pathways to ensure that the SLIPT methodology is robust under complex correlation structures and applicable to such complex genomics data.

These simulations combine the approach of prior simulation analyses (in Sections 3.3 and 6.1) with the graph structures for biological pathways (as used in Chapter 5). This enabled testing whether subtle or large differences in pathway structure affect synthetic lethal detection, whether inhibiting relationships (or inverse correlations) between genes affect synthetic lethal detection, and whether synthetic lethal detection varies by which gene is synthetic lethal and which genes are closely linked within the pathway structure. In addition, large numbers of synthetic lethal genes and biologically

feasible numbers of genes (with many non-synthetic lethal genes) was tested to replicate the findings of Sections 3.3 and 6.1 in correlated structures derived from pathway graphs, including examples of biological pathways from Reactome (Croft *et al.*, 2014).

Simple and more complex constructed graph structures were used to demonstrate the impact of pathway structure of the performance of SLIPT for synthetic lethal detection in simulations. In addition, more complex constructed graph structures were compared to the phosphoinositide 3-kinase (PI3K) and $G_{\alpha i}$ signalling pathways derived from Reactome which were used for simulation of pathway structures of biological complexity (as shown in Figures 5.1 and F.4).

6.2.1 Performance over Graph Structures

6.2.1.1 Simple Graph Structures

Simple pathway modules were used to test the effect of pathway structure on the performance of detecting synthetic lethal partners within graph structures. For an initial comparison, the graph structures (shown by Figure 6.6) were used where a gene has one upstream regulator and two downstream (Figure 6.6b) or a gene has two upstream regulators and one downstream gene (Figure 6.6b). SLIPT has a high performance in these simulations, detecting randomly selected synthetic lethal partners in both of these small simple networks (in Figures 6.7 and K.1).

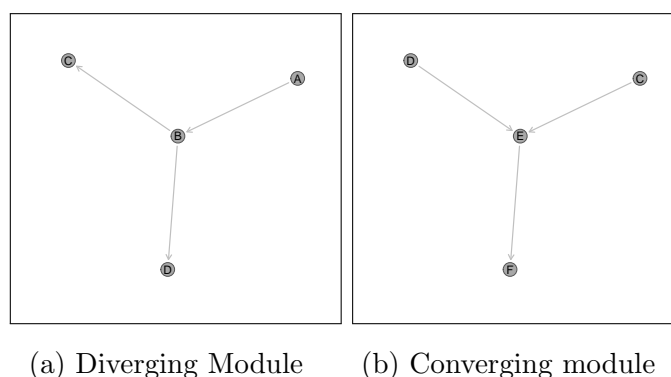
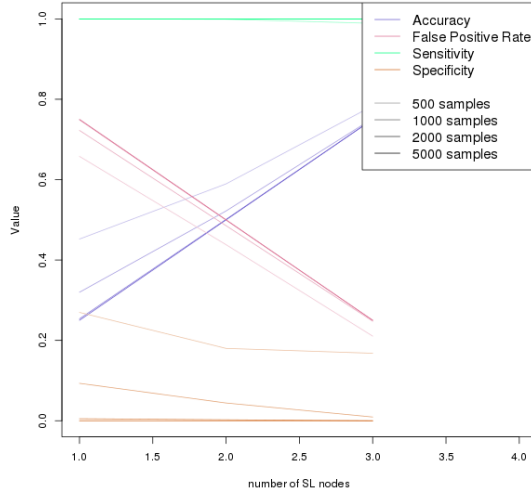
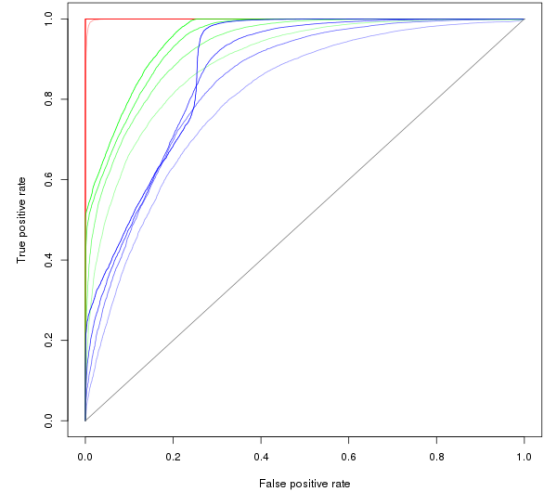


Figure 6.6: **Simple graph structures.** A simple graph structures used to demonstrate the simulation procedure. These are examples of a pathway diverging or converging respectively which enabled testing the importance of direction in pathway structures. These are used with both activating and inhibiting relationships as shown.

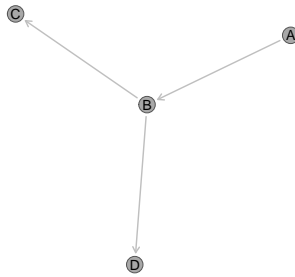
As previously observed (in Section 3.3), performance declined with higher numbers of synthetic lethal genes and lower sample sizes. However, the sensitivity of SLIPT is



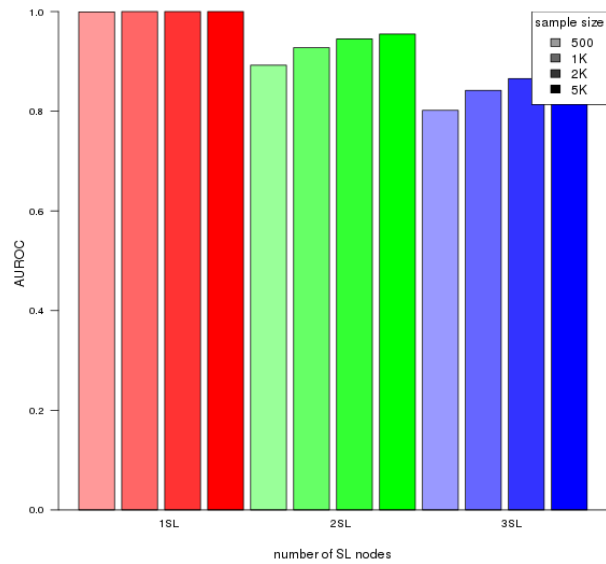
(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.7: **Performance of simulations on a simple graph.** Simulation of **synthetic lethality** was performed sampling from a multivariate normal distribution generated from a diverging **graph** structure. Performance of **SLIPT** declines for more synthetic partners but this is mitigated by increased sample sizes (in darker colours). This manifests as a decline in specificity and the false positive rate. For each parameter value, 10,000 simulations were used. Colours of the **ROC** curves in Figure 6.7b correspond to the parameters in Figure 6.7d.

high with conventional p-value thresholds (adjusted by [False discovery rate \(FDR\)](#)). [Synthetic lethal](#) partners are often distinguishable for non synthetic lethal genes, even in simple highly correlated networks. The small number of genes and their high correlation has an impact on the [ROC](#) curves for higher numbers of [synthetic lethal](#) partners which are skewed compared to those observed previously. Specificity cannot be tested if all potential partner genes are [synthetic lethal](#) which limits the number of [synthetic lethal](#) genes that can be tested.

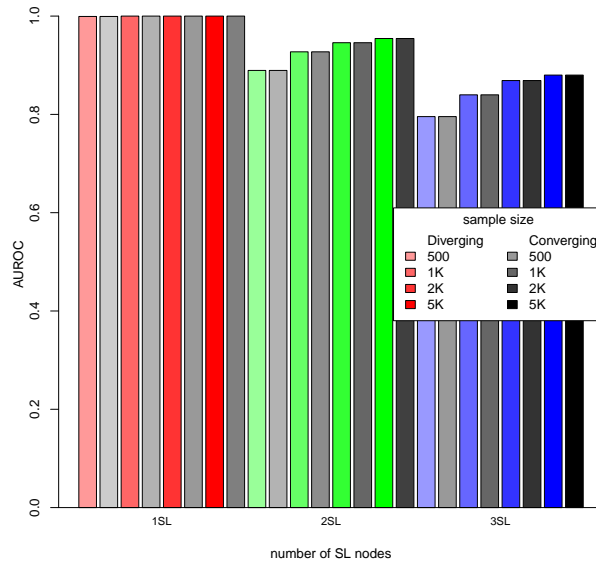
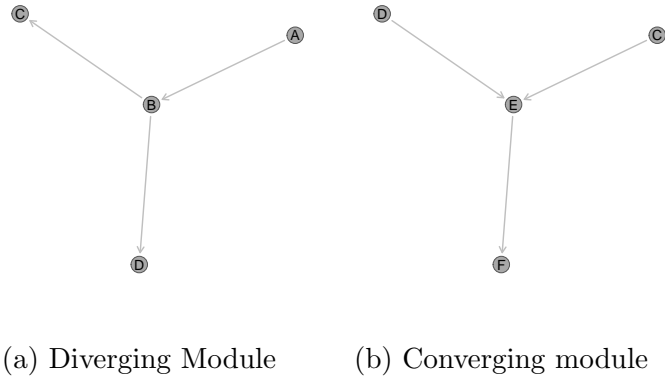
These results were consistent between the pathway modules of diverging (in [Figure 6.8a](#)) and converging signals (in [Figure 6.8b](#)). The [AUROC](#) performance and underlying curves were strikingly similar between these [graph](#) structures (as shown in [Figures 6.7](#) and [K.1](#)). Thus the performance of [SLIPT](#) was not perturbed by [pathway](#) structure, specifically the direction of pathway relationships since these [graph](#) structures also demonstrate pathways in opposite direction. In a direct comparison (shown in [Figure 6.8c](#)), the performance of simulations did not differ across parameter values in these simple [graphs](#) and therefore [SLIPT](#) is robust to pathway direction.

6.2.1.2 Constructed Graph Structures

A more complex [graph](#) structure was used to examine the performance of detecting [synthetic lethal](#) partners with [SLIPT](#) in simulated [expression](#) data with pathway correlation structures. For a simple chain of gene representing a pathway (shown in [Figure 6.9](#)), the above findings were generally replicated. Performance was high across parameter values in small networks, with similar decreases in higher numbers of [synthetic lethal](#) genes to detect and lower sample size.

When detecting [synthetic lethal](#) genes with [SLIPT](#) using adjusted ([FDR](#)) p-value thresholds, the performance differences were primarily due to changes in specificity, as the small numbers of [synthetic lethal](#) genes still had highly significant p-values. Despite lower specificity and performance in [ROC](#) curves, the accuracy increases and false positive rate decreases desirably with higher numbers of [synthetic lethal](#) genes due to the high sensitivity and proportion of [synthetic lethal](#) genes detected. Therefore the thresholds imposed by adjusted p-values appear to be appropriate for detecting [synthetic lethal](#) partners, even in strongly correlated pathways, at least in these small-scale test cases.

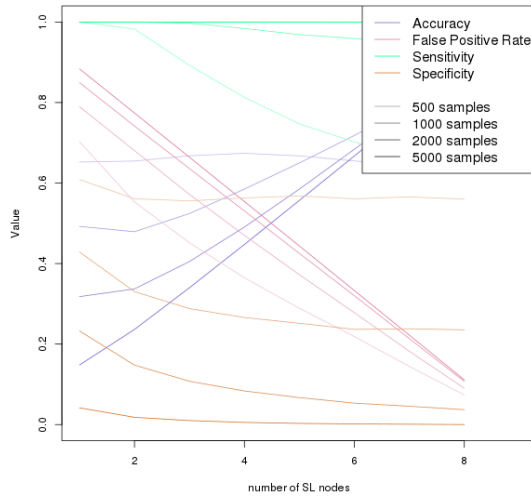
An artifact of these small test cases led to the skewed [ROC](#) curves (as discussed in [Section 6.2.1.1](#)), which may be the result of the low number of non-synthetic lethal genes to identify as true negatives, affecting the accuracy of specificity. This issue does not occur in larger, more complex [graph](#) structures, even with modest total numbers



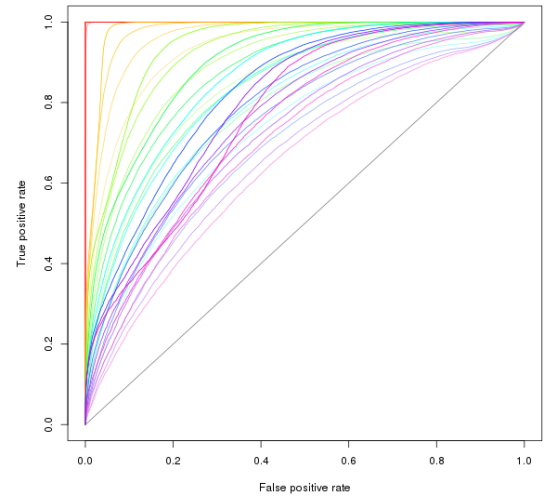
(c) Performance between Graph Structures

Figure 6.8: **Performance of simulations is similar in simple graphs.** The AUROC values for simulations of multivariate normal distributions based on each graph structure yielded indistinguishable performance across parameter values in 10,000 simulations.

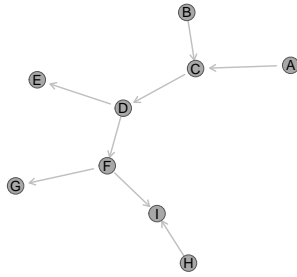
of genes and high correlations (as shown in Section 6.3). This issue is unlikely to occur in large expression datasets with many non synthetic lethal genes, as shown previously (in Section 3.3 and 6.2.1.1) with graph structures in larger datasets (in Section 6.2.4).



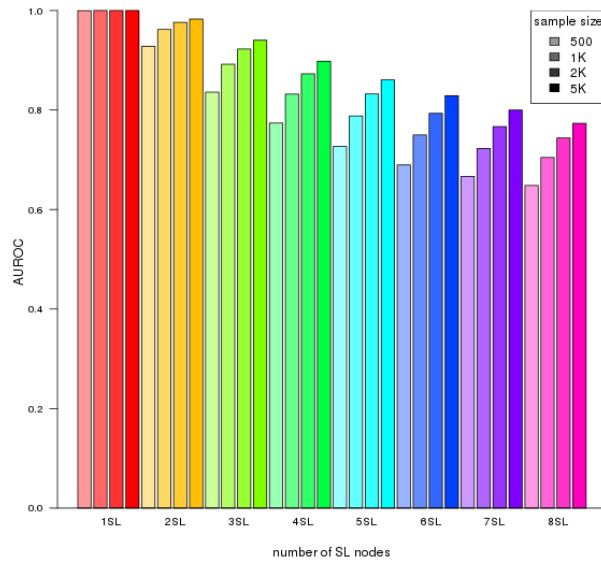
(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.9: **Performance of simulations on a pathway.** Simulation of [synthetic lethality](#) was performed sampling from a multivariate normal distribution generated from a [pathway](#) structure. Performance of [SLIPT](#) declines for more synthetic partners and lower sample sizes (in darker colours). For each parameter value, 10,000 simulations were used. Colours of the [ROC](#) curves in Figure 6.9b correspond to the parameters in Figure 6.9d.

6.2.2 Performance with Inhibitions

Simulations of [synthetic lethality](#) in [expression](#) data were also performed with correlation structures derived from [graphs](#) containing inhibiting relationships (as are commonplace in biological pathways) which produce negative correlations. As shown in Figure 6.10, these are not an issue for detection by [SLIPT](#). Rather, the [SLIPT](#) procedure performs well on simple graph modules with highly negative correlations. With [synthetic lethal](#) detection based on p-value (adjusted by [FDR](#)), there was higher specificity, higher accuracy, and lower false positive rate in an inhibitory graph than the same graph with activating relationships (as shown by Figure 6.7).

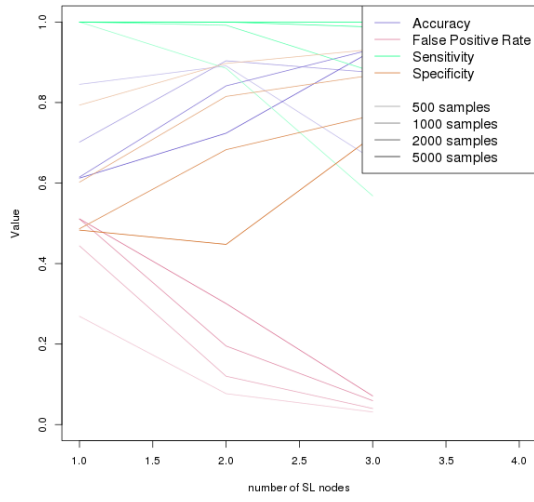
The [ROC](#) curves for an inhibiting graph also show consistently high specificity irrespective of detection threshold with only the upper extreme of the curve exhibiting a skew below random performance (in Figure 6.10). Nevertheless, the [AUROC](#) values show a high performance across parameter values, particularly avoiding issues with higher numbers of [synthetic lethal](#) partners (as observed in Section 6.2.1.1). However, performance was marginally lower for higher numbers of [synthetic lethal](#) genes to detect and lower sample sizes, consistent with previously observations.

Negatively correlated simulated datasets are also unperturbed by minor differences in [graph](#) structure, such as changing in the direction of the graph module. As observed for activating relationships in these graph modules, the performance was highly concordant between the graph modules (shown by similar results in Figures 6.10 and K.2).

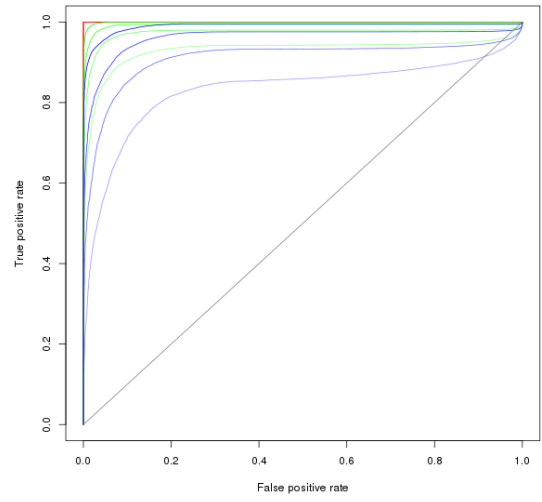
Detection of [synthetic lethality](#) by [SLIPT](#) in simulated data with inhibiting relationships outperforms simulations with activating relationships in the same [graph](#) structure (as shown in Figure 6.11). Thus [SLIPT](#) is robust in [gene expression](#) datasets with inverse correlations and performs well in them, at least in simple test cases. This is important because such relationships occur frequently in biological pathways and therefore the findings inferred from [graph](#) structures without inhibiting relationships are a conservative estimate.

The [SLIPT](#) methodology likely performs better in biological pathways (which contain negative correlations) than the [graph](#) structures discussed previously (in Section 6.2.1). This is likely since negative correlations lead to [synthetic lethal](#) partners and inversely correlated genes which are positively correlated with the query gene. As previously shown, the [SLIPT](#) methodology performs well with specificity against positively correlated query genes (in Sections 3.3.2.2 and 6.1.2.1).

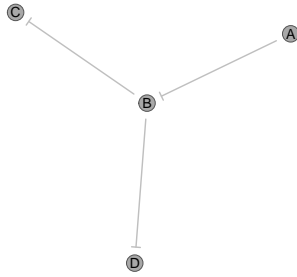
Similarly, more complex [graph](#) structures with entirely inhibiting relationships (negative correlations) also perform desirably on p-value thresholds (adjusted by [FDR](#)) and



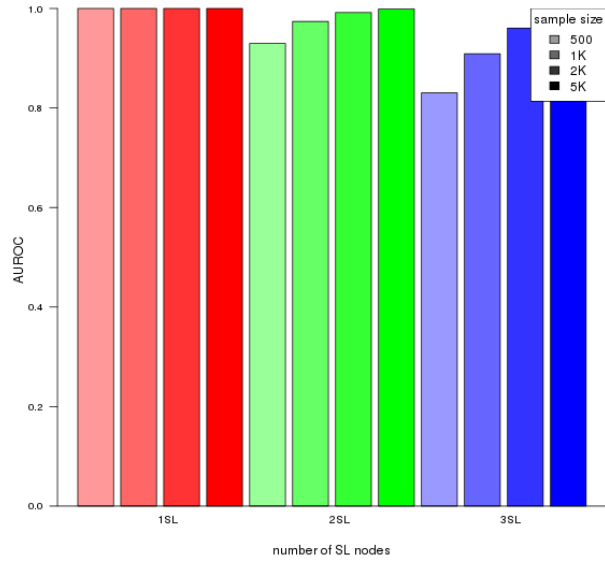
(a) Statistical evaluation



(b) Receiver operating characteristic

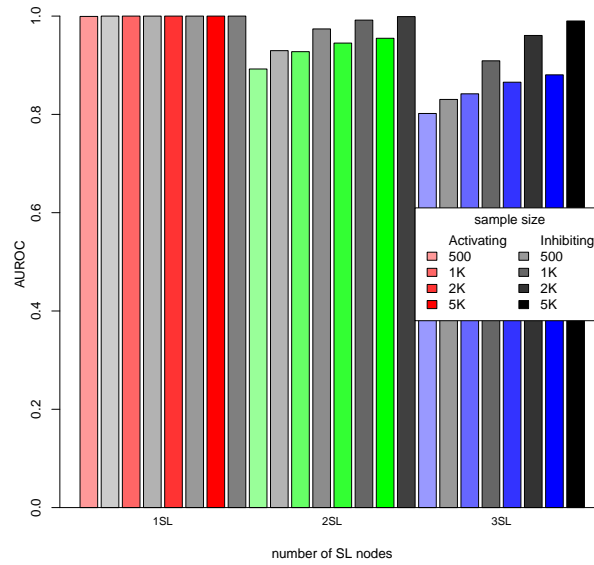
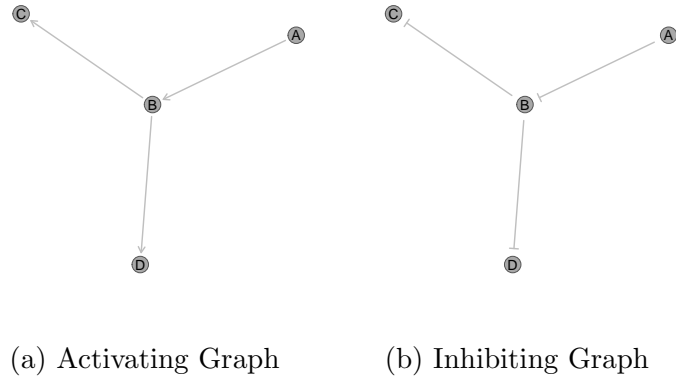


(c) Graph Structure



(d) Statistical performance

Figure 6.10: **Performance of simulations on a simple graph with inhibition.** Simulation of [synthetic lethality](#) was performed sampling from a multivariate normal distribution generated from an inhibiting graph. Performance of [SLIPT](#) declines for more synthetic partners and lower sample sizes. For each parameter value, 10,000 simulations were used. Colours of the [ROC](#) curves in Figure 6.10b correspond to the parameters in Figure 6.10d.

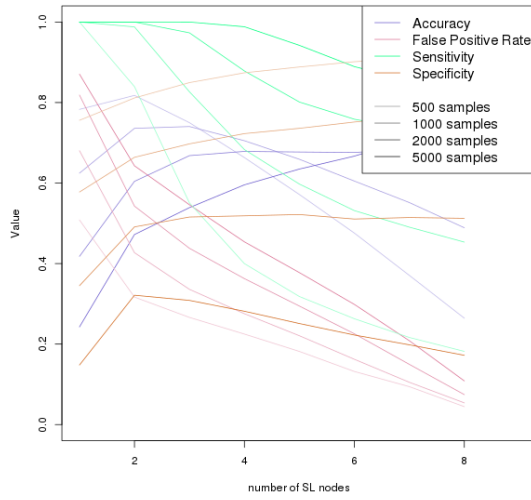


(c) Performance between Graph Structures

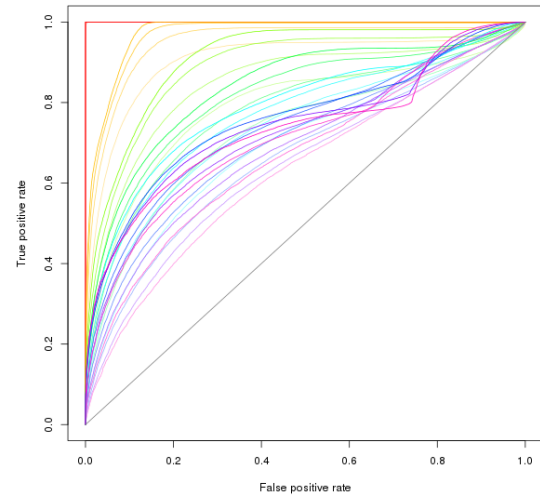
Figure 6.11: **Performance is higher on a simple inhibiting graph.** The AUROC values for simulations of multivariate normal distributions based on inhibitions in the Graph structure yielded consistently higher performance across parameter values in 10,000 simulations.

have high performance across increasing numbers of **synthetic lethal** genes, particularly for sufficiently high sample sizes (as shown by Figure K.3). However, this is not necessarily the case for **graph** structures with a combination of activating and inhibiting relationships (i.e., containing positive and negative correlations) As shown by Figure K.4, such a mixed **network** structure does not necessarily have high performance across parameters as observed for purely inhibiting networks.

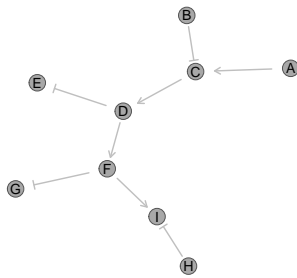
These still appear to have desirably high sensitivity, high accuracy, and low false positive rate for detecting more **synthetic lethal** genes, despite poor specificity. The



(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.12: **Performance of simulations on a constructed graph with inhibition.** Simulation of **synthetic lethality** was performed sampling from a multivariate normal distribution generated from **pathway** structure with a combination of inhibitions. Performance of **SLIPT** declines for more synthetic partners and lower sample sizes. For each parameter value, 10,000 simulations were used.

ROC curves are particularly skewed for high proportions of the network being [synthetic lethal](#) and may stem from low numbers of true negative genes to detect (as discussed in Section 6.2.1.1). In a direct comparison of performance (shown in Figure 6.13), the purely inhibiting graph had consistently higher performance than the activating one as observed for simpler [graphs](#) (in Figure 6.11).

In contrast, the combination of activating and inhibiting relationships had slightly lower performance across parameters compared to the same [graph](#) structure with activating relationships. Therefore correlation structure can impact on the performance of SLIPT in a graph network, in either direction, specifically the addition of negative correlations. However, this may be an artifact of the simulation procedure as [synthetic lethal](#) genes from the correlation structure were randomly selected (without regard to their relationships), with the query gene added to ensure that conditions for [synthetic lethal](#) relationships were met.

This system for simulating inhibitory pathways is not ideal since it lead to [synthetic lethal](#) gene combinations, by randomly selecting them, which are unlikely to occur in biological pathways. These randomly selected [synthetic lethal](#) genes may account for the detection results being suboptimal (i.e., difficult to detect [synthetic lethal](#) partners) compared to previous investigations. It is expected that inversely correlated synthetic partner genes will be highly expressed in a mutually exclusive manner such that at least one of them will be compensating for loss of the query gene in most samples, leading to a weak [synthetic lethal](#) signature in [expression](#) data in this case. Furthermore, this case may not be representative of empirical biological data with [synthetic lethal](#) partners of [tumour suppressor](#) genes which are commonly inversely correlated to the query gene (to some extent) and therefore it is unlikely that they are strongly negative correlated with each other, unless they are [synthetic lethal](#) partners of each other as well. It is plausible that many [synthetic lethal](#) partner genes will serve to separately compensate for the loss of query gene function and be positively correlated with each other. Nonetheless, these simulations are sufficient to demonstrate that correlation structure (particularly negative correlations) have an impact on the detection of [synthetic lethality](#). However, SLIPT is still able to perform well across [graphs](#) with different activating and inhibiting relationships and the perturbations in performance are marginal, particularly those reducing performance compared to an activating network.

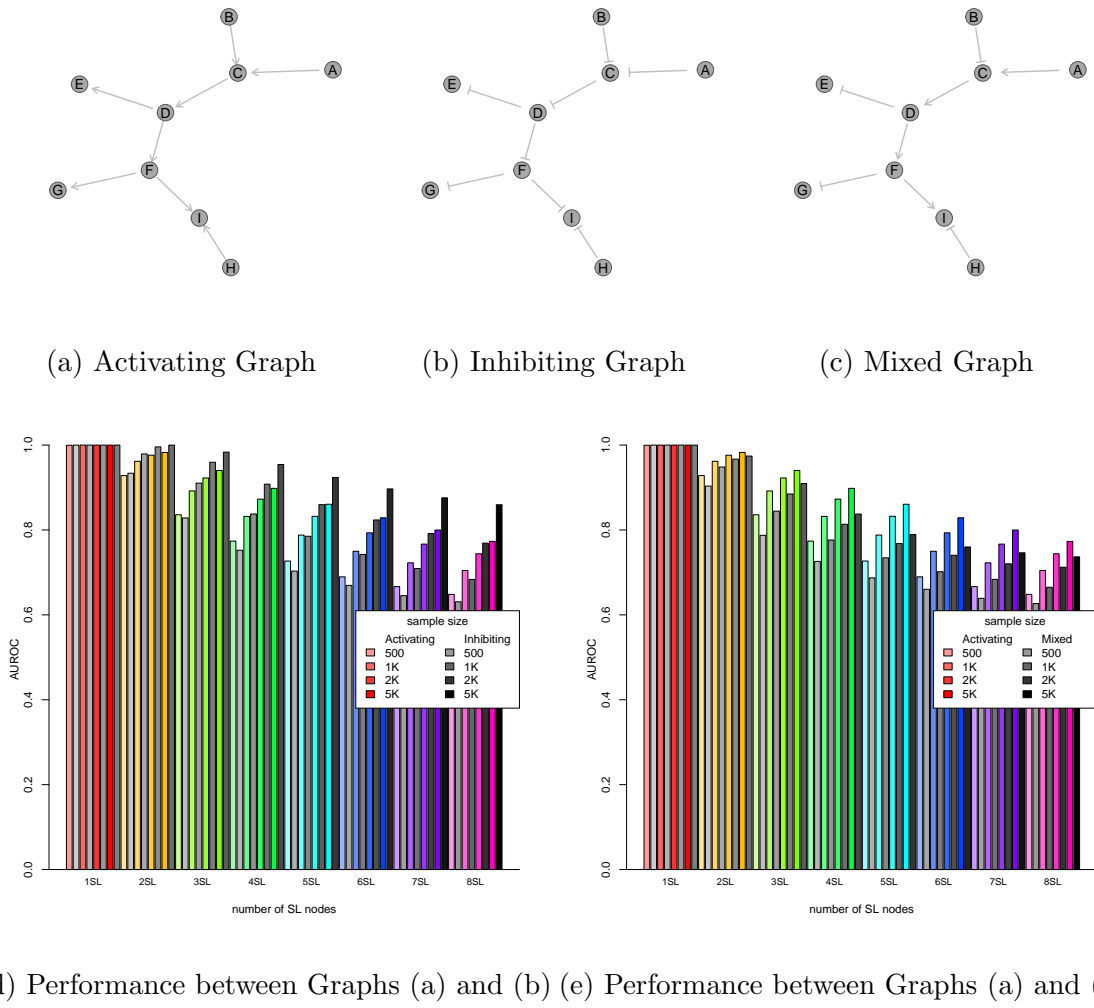


Figure 6.13: **Performance is affected by inhibition in graphs.** The AUROC values for simulations of multivariate normal distributions based on graph structure containing only inhibitions in the Graph structure yielded consistently higher performance across parameter values in 10,000 simulations. A combination of activating and inhibiting relationships had lower performance but was more similar to the activating graph.

6.2.3 Synthetic Lethality across Graph Structures

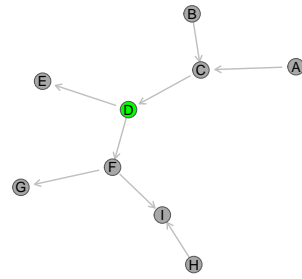
While synthetic lethal genes are distinguishable in principle from those highly positively correlated with them (as shown by ROC analysis), they are not necessarily distinguished as reflected by low specificity and high false positive rates in poorly performing simulations throughout this section. The negative correlations are not subject to the same issue, they sometimes perturb the correlation structure between synthetic lethal partner genes making it difficult to detect many of them. Thus far, synthetic

lethal genes have been selected randomly which is a limited approach. To examine the impact of pathway relationships in more detail, specific genes will be selected to be synthetic lethal within a network. Replicate simulations were performed for synthetic lethal detection with a fixed synthetic lethal gene, in contrast to previous investigations (randomly selecting synthetic lethal genes). This investigation was performed to demonstrate the impact of these genes being synthetic lethal in the detection of neighbouring genes in the pathway network, under graph structure activating and inhibiting relationships.

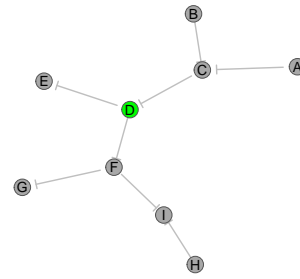
For instance, detection of a synthetic lethal gene in an activating graph structure (as shown in Figure 6.14a) is straightforward: the χ^2 values across simulations are clearly distinguishable from non synthetic lethal genes (shown in Figure 6.14c). A small number of simulations were performed for each gene being designated as synthetic lethal. In each case (of each gene being the synthetic lethal partner), the synthetic lethal gene was detectable with highest χ^2 value, being distinguishable amongst 20,000 genes including the highly correlated graph network (as shown in Figure K.5).

This is consistent with previous observations that SLIPT performed optimally for a single synthetic lethal partner in this network (in Figure 6.9). Despite optimal performance in a ROC curve irrespective of detection threshold, many of the highly correlated genes would be detected as false positives using a conventional p-value threshold (even if adjusted by FDR) from a χ^2 test with 4 degrees of freedom as performed by SLIPT (as described in Section 3.1). In particular, the genes that are adjacent to the synthetic lethal gene “D” within the graph structure exhibited high test statistics across simulations which would often be reported as false positives (as shown in Figure 6.14c). This is not specific to example of gene “D”, with the neighbouring genes exhibiting higher χ^2 test statistics for each gene in the network when it is designated as the synthetic lethal partner (as shown in Figure K.5).

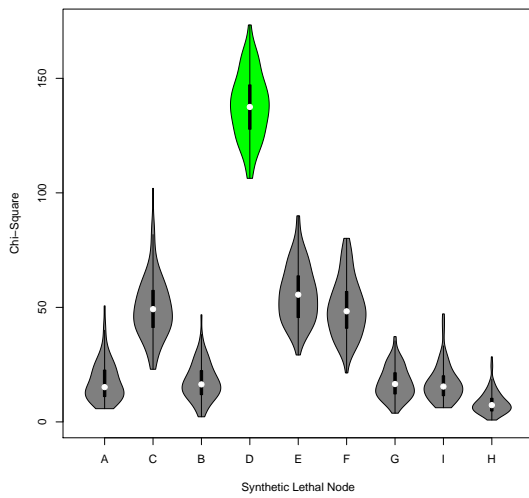
Thus the synthetic lethal signal propagates from the true synthetic lethal gene throughout the network such genes nearer to the true synthetic lethal gene (more highly correlated) have higher test statistics and are more likely to be detected by SLIPT as false positives. This tendency for adjacent genes to be detected as synthetic lethal false positives is consistent with the synthetic lethal pathways being more concordant between SLIPT in TCGA data (TCGA, 2012) and the siRNA screen (Telford *et al.*, 2015) than individual gene results (in Chapter 4). False positive genes are therefore still more likely to be involved in a synthetic lethal pathway by being correlated with a true synthetic lethal gene and synthetic lethal pathways are likely to have



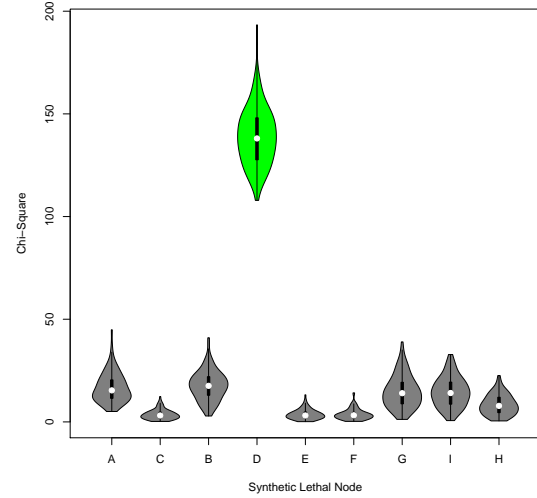
(a) Activating Graph



(b) Inhibiting Graph



(c) χ^2 distribution for Activating Graph



(d) χ^2 distribution for Inhibiting Graph

Figure 6.14: **Detection of synthetic lethality within a graph structure.** The gene “D” was designated to be **synthetic lethal** and the χ^2 value from **SLIPT** was computed for each gene across each **graph** structure. The χ^2 values were computed in 100 simulations of datasets of 20,000 genes including the **graph** structure and 1000 samples. Adjacent genes exhibited lower χ^2 values with inhibiting relationships.

many genes detected by **SLIPT** giving a consensus of evidence, supporting the pathway over-representation approach in particular which may account for how it differs from pathway **metagenes**. Furthermore, **SLIPT** is still viable to detect true **synthetic lethal** partners or prioritise those most likely to be experimentally validated since those with the strongest support (i.e, higher χ^2 values and more significant p-values) are more likely to be the underlying **synthetic lethal** gene.

In contrast to an activating graph (Figure 6.14a), the immediately adjacent genes in an inhibiting graph (Figure 6.14b) had neither an elevated χ^2 test statistics indicating **synthetic lethality** nor a significant inverse effect (as shown in Figure 6.14d). Similar simulations were performed a **graph** structure with inhibiting relationships within a dataset of 20,000 genes. The adjacent genes to the **synthetic lethal** gene “D” did not have elevated χ^2 values and therefore true **synthetic lethal** partners were highly distinguishable from non synthetic lethal genes with inhibiting relationships. This was not specific to “D” and was shown across any gene in the **graph** structure if it were designated to be the **synthetic lethal** partner of the query gene (shown in Figure K.6). This is consistent with the detection of many genes involved in kinase signalling, gene regulation, and other known cancer pathways (in Chapter 4) which frequently have inhibitory steps. These results support **SLIPT** as an appropriate approach to distinguish **synthetic lethal** partners in biological pathways, including those relevant to cancer growth and inhibition.

However, it should be noted that the 2nd degree neighbours of the **synthetic lethal** gene still exhibited moderate χ^2 values (and are moderately correlated with the **synthetic lethal** gene). It is still possible for these to be detected as false positives as previously described for an activating **graph** structure although the presence of inhibitory relationships (and negative correlations) further increases the differences in test statistics for correlated genes and underlying **synthetic lethal** partners as shown by the extreme example (in Figure K.6).

These findings are consistent with simulations in a graph containing a combination of activating and inhibiting relationships which exhibits a either of these χ^2 profiles depending on which gene is **synthetic lethal** and the relationships to adjacent genes (as shown in Figure K.7). Note that in this case, the **synthetic lethal** gene is distinguishable and inhibitory relationships within this **graph** structure make it easier to detect underlying **synthetic lethal** genes with **SLIPT** by a more highly significant χ^2 test. This contrasts with randomly selecting multiple **synthetic lethal** genes (in Figure 6.13) where the performance of **SLIPT** was impeded by the inhibitory relationships between **synthetic lethal** partners in this **graph** structure. Therefore the random **synthetic lethal** genes selected previously with negative correlations between them which had poor performance are likely to have created an artifact in the simulation results as they are biologically implausible and constrain the **synthetic lethal** simulation procedure

The results with one **synthetic lethal** partner were sufficient to infer the impact of **synthetic lethal** partners within pathways on neighbouring (correlated) genes. However,

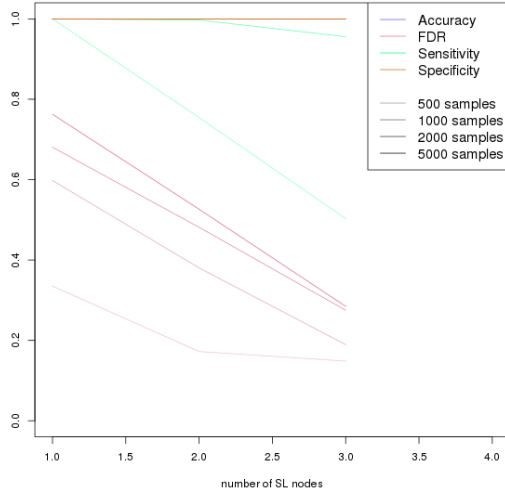
it is plausible that the [synthetic lethal](#) signatures in [expression](#) data would propagate through a network with multiple [synthetic lethal](#) partners as sources, provided that the correlations between [synthetic lethal](#) partners is biological feasible. These simulations were performed on a correlated [graph](#) structure within a larger [gene expression](#) dataset of 20,000 genes (as performed in Sections 3.3 and 6.2.4), a feasible number for a full human [gene expression](#) dataset, and as such are comparable to the findings below.

6.2.4 Performance within a Simulated Human Genome

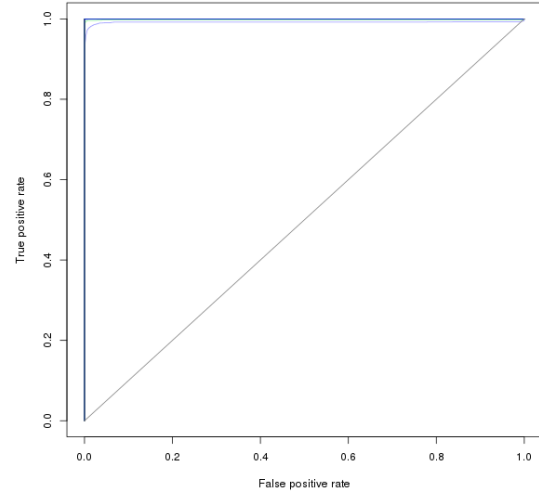
The high proportion of [synthetic lethal](#) partners in small networks (and fewer true negatives) made it difficult to accurately assess the performance of SLIPT with higher numbers of true partners to detect (as noted in Section 6.2.1.1). The addition of more non synthetic lethal genes in previous simulations increased the performance of SLIPT, particularly the specificity to reduce the number of false positives (as shown in Sections 3.3 and 6.1). The [graph](#) structures (as used in Section 6.2.1) of genes with correlations from sampling a multivariate normal distribution were included in a larger simulated dataset of 20,000 genes. This simulation procedure enabled testing of the performance of SLIPT when detecting [synthetic lethal](#) partners within correlated [graph](#) structures (of a [synthetic lethal](#) pathway) in the context of biologically feasible numbers of genes.

The simulations performed in Section 6.2.1.1 were replicated within a dataset of 20,000 genes with the rest being composed on non synthetic lethal genes without correlation structure. The aforementioned issue with specificity in a higher number of underlying [synthetic lethal](#) genes did not occur in a simple [graph](#) structure (as shown in Figure 6.15). For such a small graph module of highly correlated genes within a [gene expression](#) dataset, detection of [synthetic lethal](#) genes within the network by SLIPT and distinguishing these from the larger dataset had high performance across parameter values. In this case, a reduction in sensitivity resulted in poorer performance, a higher number of non synthetic lethal genes were correctly identified, with a low false positive rate and high accuracy. Thus the use of stringent χ^2 p-value thresholds (adjusted by FDR) are suitable for testing for [synthetic lethality](#) in [gene expression](#) data across the number of genes in human and cancer data.

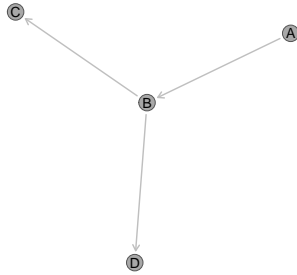
In a direct comparison with simulations in the [graph](#) structure alone (as performed in Section 6.2.1.1), detection of [synthetic lethality](#) with SLIPT performs consistently better in a larger dataset with many true negative genes to detect (as shown in Figure 6.16). The SLIPT methodology as it has a high specificity and low false positive



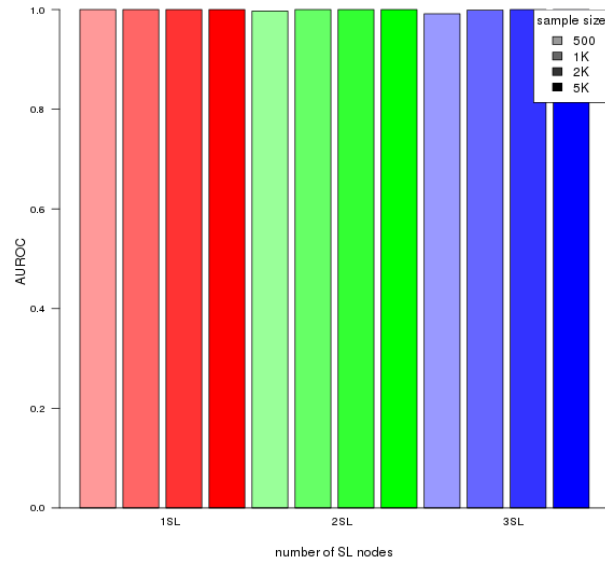
(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.15: **Performance of simulations including a simple graph.** Simulation of **synthetic lethality** was performed sampling from a multivariate normal distribution (without correlation structure apart from the graph shown). Performance of **SLIPT** was high across parameters for detecting **synthetic lethality** in the **graph** structure within a larger dataset. The sensitivity decreases for a greater number of true positives to detect but the specificity remains high with a low false positive rate.

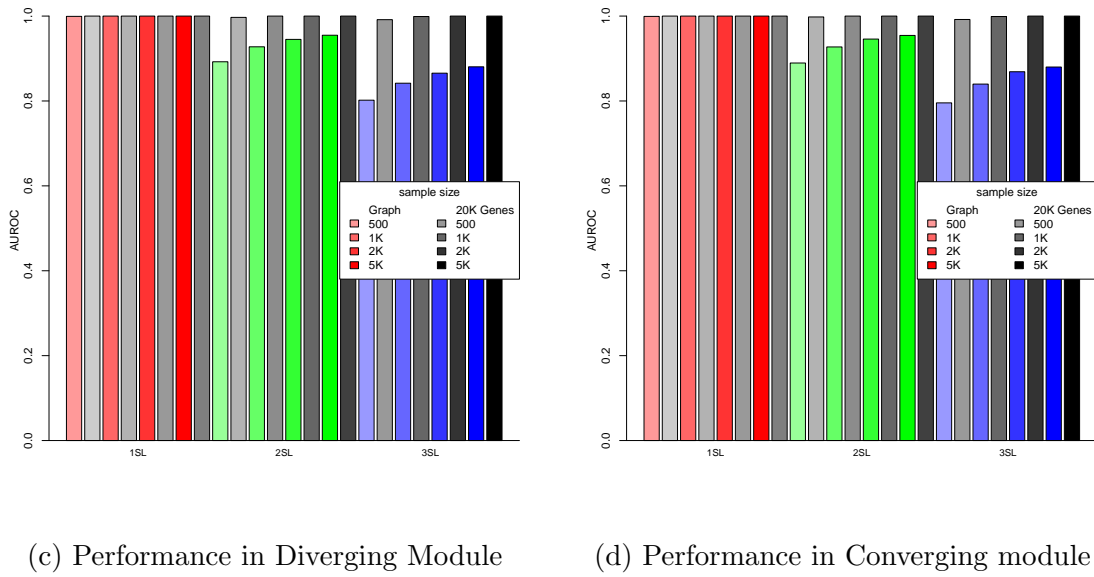
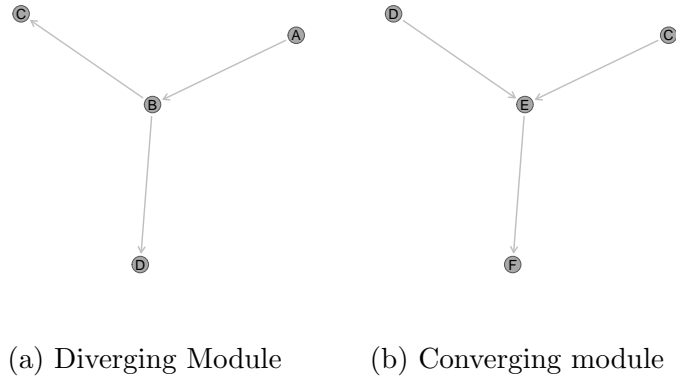


Figure 6.16: **Performance on a simple graph improves with more genes.** Simulations were performed with each of the [graph](#) structures to detect [synthetic lethal](#) partners within them. In either structure, performance of detection in a dataset containing on the [graph](#) structure (in colour) was lower than testing the [graph](#) structure within a larger dataset of non synthetic lethal genes (without correlations).

rate, which is desirable. [SLIPT](#) is therefore applicable to large [gene expression](#) datasets where these are important considerations since the number of negative genes to correctly identify often vastly outnumbers the number of positive genes to detect.

This increase in performance with more non synthetic lethal genes to detect does not necessarily apply in an inhibiting [graph](#) structure. This an important consideration since biological pathways commonly contain inhibiting relationships (and inverse correlations), although they are unlikely to occur across an entire pathway. While an

increased performance for an activating graph was replicated, in this case, the performance of simulations of an entirely inhibiting graph structure did not improve within a larger dataset. Rather, the performance in the inhibiting graph structure was similar to simulations of the graph structure in isolation. It is expected that the findings based on these simulations of genes with pathway structures in smaller datasets (as described in Section 6.2.1) will be relevant to larger datasets. The simulation results in these inhibiting graph structures perform comparably or higher with more non synthetic lethal genes to distinguish from them even with inhibitory relationships within the graph structure

Performance of synthetic lethal detection of SLIPT in graph structures with inhibitions included in a larger dataset of non synthetic lethal genes did not diminish to the level of the graph structure simulated alone. In some cases (as shown in Figure 6.17), the performance of an inhibitory graph structure was consistently elevated when included within a larger data. However, these did not perform as well as the equivalent graph structures without inhibitory relationships within a similar dataset.

This poorer performance of inhibitory graph structures may be due to highly negatively correlated genes being false positives. These genes will be positively correlated with the query gene if they are negatively correlated with a synthetic lethal partner (i.e., within a synthetic lethal pathway). The SLIPT procedure performs well at distinguishing these positively correlated genes, as previously shown (in Sections 3.3.2.2 and 6.1.1.1). These false positives will also be a minority amongst a larger dataset of non synthetic lethal genes without correlation to the query or synthetic lethal genes.

It more likely that the poorer performance stems from negative correlations between synthetic lethal genes which makes them more difficult to individually detect (as observed in Section 6.2.2). As discussed in Section 6.2.3, this is likely an artifact of the simulation procedure selecting random synthetic lethal genes with strong inhibitory relationships between them. Therefore the poorer performance for inhibiting graphs within larger datasets is not cause for concern because the cases where SLIPT performs poorly are likely to be combinations of simulated synthetic lethal genes that are not likely to occur within biological pathways. This simulation procedure has included higher-order synthetic lethal to produce the weakest signal of synthetic lethality for individual partner genes which are still detectable by SLIPT.

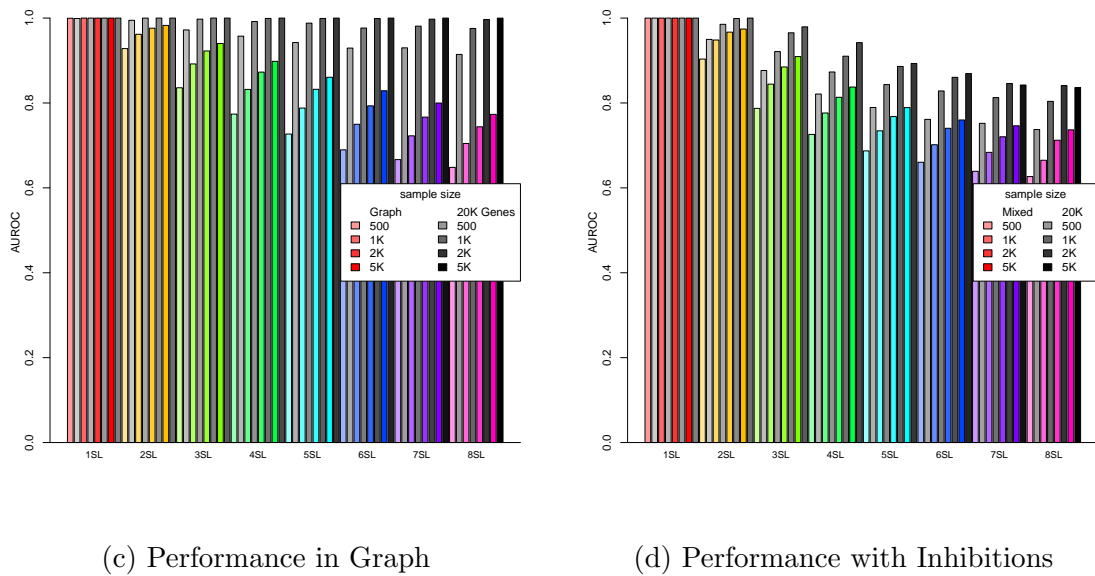
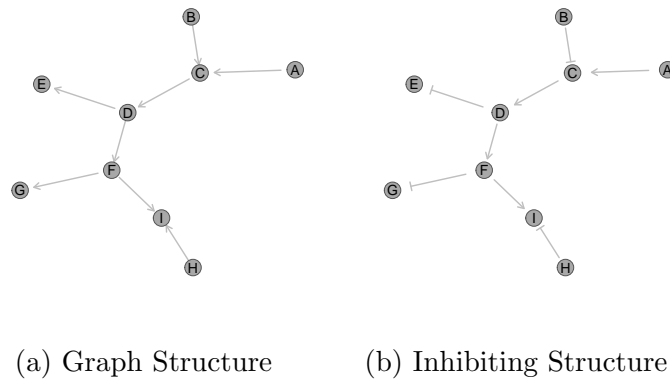


Figure 6.17: **Performance on an inhibiting graph improves with more genes.** Simulations were performed in a [graph](#) structure with activating and inhibiting relationships to detect [synthetic lethal](#) partners within them. In contrast to an activating graph, performance of detection in a dataset containing only the [graph](#) structure (in colour) was as much lower than testing the [graph](#) structure within a larger dataset of non synthetic lethal genes (without correlations) in an inhibiting [graph](#) structure with negative correlations.

6.3 Simulations in More Complex Graph Structures

Investigations with simulations based on [graph](#) structures were extended to larger [graphs](#), enabling more synthetic lethal genes within a pathway and modelling the complexity of a biological pathway. Sensitivity declines over a greater range for the number of [synthetic lethal](#) partners in a larger network with a trade-off with specificity (as shown in Figures [K.8–K.10](#)). However, the accuracy declined for greater numbers of

[synthetic lethal](#) partners and the false positive rate peaks at intermediate values. In this range, difference between simulations varied with greater sample size. The [AUROC](#) results were similar between these more complex [graph](#) structures, , although the larger [graph](#) (Figure [K.10](#)) differed in sensitivity and specificity at an adjusted (FDR) p-value threshold. This difference be due to different proportions of [synthetic lethal](#) and non-synthetic lethal genes to detect, since these [graphs](#) (in Figures [K.8](#) and [K.9](#)) had fewer genes.

While the [graph](#) structures (of similar size) were highly distinct, they had similar performance profiles across parameters. [SLIPT](#) is therefore robust across [pathway](#) structures and is more affected by the number or proportion of genes to detect. Findings from previous simulations in similar correlation structures (in Section [3.3](#)) should be applicable to [expression](#) data with more complex correlation structures, such as those occurring in biological pathways. Specifically, [synthetic lethal](#) partners are distinguishable from closely correlated genes in the context of a biological pathway networks, irrespective of thresholds (shown by [ROC](#)) and with the sensitivity and specificity of p-value thresholds (adjusted by [FDR](#)) as used for [SLIPT](#) (in Chapters [4](#) and [5](#)).

The findings for inhibitory [graph](#) structures were replicated with larger more complex [graph](#) structures with inhibiting relationships and more [synthetic lethal](#) genes to detect (shown in Figures [K.11–K.14](#)). In each [graph](#) structure, simulations entirely with inhibiting relationships (Figures [K.11](#), [K.13](#), and [K.15](#)) had higher performance than the equivalent graph with entirely activating relationships (Figures [K.8](#), [K.9](#), and [K.10](#)) or a combination of activating and inhibiting relationships (Figures [K.12](#), [K.14](#), and [K.16](#)). While the presence of negative correlations subtly affects the performance of [SLIPT](#), the methodology is robust across the exact structures of genes and is therefore applicable to detecting [synthetic lethal](#) genes in a range of (synthetic lethal) biological pathways with different structural relationships.

6.3.1 Simulations over Pathway-based Graphs

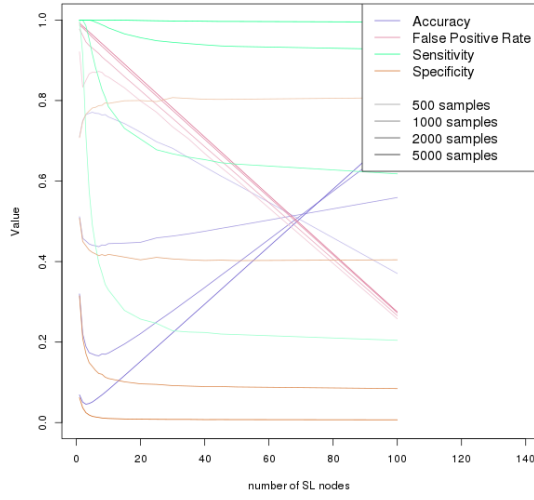
Simulations of [synthetic lethality](#) in [gene expression](#) with correlation structures thus far have used simple blocks of correlated genes (as used in Section [3.3](#)) or derived from constructed [graph](#) structures (as used in Section [6.2](#)). These have been used to make inferences on the impact of correlation structure, it remains to be shown whether these findings are reproducible in the complexity of the biological [network](#) structure. Specifically, [SLIPT](#) was tested on simulated data with known underlying simulated

[synthetic lethal](#) partners (as described in Section 3.2.2) with multivariate normal correlation structure derived from biological pathways (as described in Section 3.4.2).

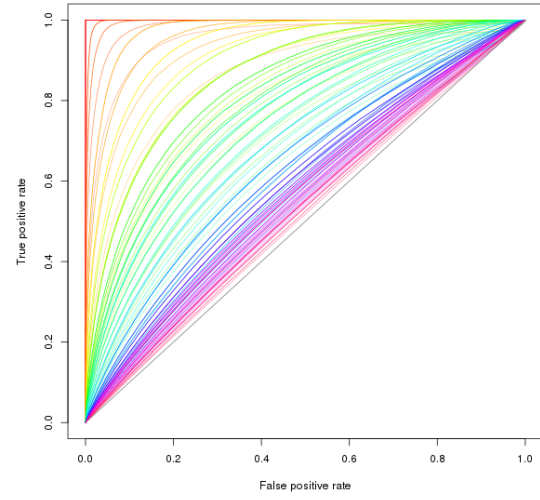
The Reactome [pathway](#) structure for the [PI3K](#) cascade (as used in Chapter 5) was used to demonstrate the simulation procedure for detecting [synthetic lethality](#) in the [graph](#) structure of a biological pathway. This pathway has clear directionality and related signalling pathways were among those identified to be [synthetic lethal](#) candidates (in Chapter 4). The [PI3K](#) pathway has a relatively moderate size (138 genes) and complexity. It is therefore suitable for comparison to previous [graph](#) structures of a similar scale (50–100 genes) with the complexity of a biological pathway.

The performance of [synthetic lethal](#) detection with [SLIPT](#), in simulated [expression](#) data based on the Reactome [PI3K](#) pathway (as shown in Figure 6.18), was concordant with previous findings. [SLIPT](#) had high performance when detecting a low number of [synthetic lethal](#) genes which decreased for high numbers of [synthetic lethal](#) genes or lower sample sizes. In particular, the performance of simulations in the [PI3K](#) pathway highly resembled the simulation results for constructed [graphs](#) of similar scale and complexity (as shown in Figures K.8 and K.9). Using thresholds based on the χ^2 p-value (adjusted by [FDR](#)), simulations in the biological [PI3K](#) pathway had a higher sensitivity and lower specificity. While the performance decreased for more [synthetic lethal](#) genes to detect within the simulated [PI3K](#) pathway, this primarily involved a reduction in sensitivity to detect [synthetic lethal](#) genes rather than false positives, as the false positive rate decreased, the accuracy increased, and the specificity was relatively unperturbed (being more dependent on sample size). Thus [SLIPT](#) was stringent in an example of a biological [graph](#) structure and is appropriate for detection of [synthetic lethal](#) genes in complex correlation structures in [gene expression](#) data involving biological pathways.

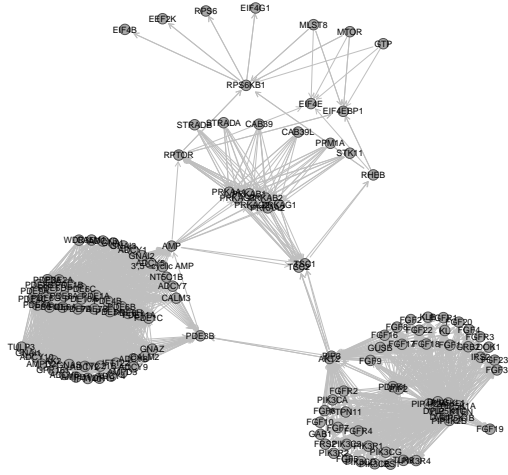
These simulations were replicated in the more complex $G_{\alpha i}$ signalling pathway (of 292 genes), which was one of the most well supported [synthetic lethal](#) pathways with loss of *CDH1* in cancer (in Chapters 4 and 5). This pathway showed similar relationships between sensitivity, specificity, and false positive rate with number of [synthetic lethal](#) partners and sample size (as shown in Figure K.17). While the overall performance was lower than for smaller networks structures, many of the findings from previous networks were replicated in a larger more complex biological network. In the $G_{\alpha i}$ signalling pathway, [SLIPT](#) had high performance for detecting low numbers of [synthetic lethal](#) genes and was highly stringent against false positives for higher numbers of [synthetic lethal](#) genes.



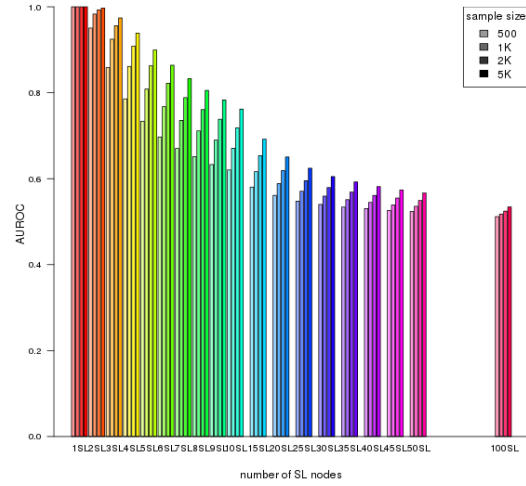
(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.18: **Performance of simulations on the PI3K cascade.** Simulation of **synthetic lethality** was performed sampling from a multivariate normal distribution based on the Reactome PI3K cascade. Performance of **SLIPT** was high across parameters for detecting **synthetic lethality** in the **graph** structure within a larger dataset. The performance decreases for a greater number of true positives to detect but the accuracy increases with a low false positive rate.

6.3.2 Pathway Structures in a Simulated Human Genome

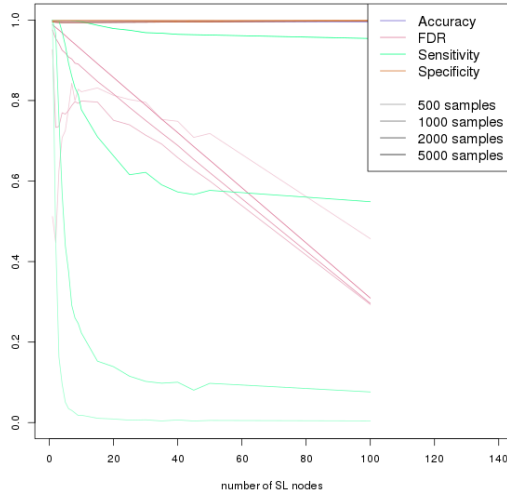
Simulations were also performed with [graph](#) structures from biological pathways included in a larger dataset to simulate [gene expression](#) data of the scale typical for human and cancer studies. These simulations (as discussed in Section 6.2.4) showed a higher specificity and therefore [SLIPT](#) had higher performance. The simulated [PI3K](#) pathway (as shown in Figure 6.19) is no exception, with high performance across parameter values, which remained high up to many genes. While the sensitivity decreases for high numbers of [synthetic lethal](#) genes to detect within the [PI3K](#) pathway, the [SLIPT](#) methodology remains accurate with high specificity in a large simulated [gene expression](#) dataset.

Therefore the [SLIPT](#) methodology is a highly stringent approach suitable to be applied for detecting [synthetic lethal](#) genes and pathways within highly complex [expression](#) data with biological [pathway](#) structure. Even the poorly performing simulations were highly stringent with low false positive rates, which are an important consideration in a [gene expression](#) data with many non synthetic lethal genes. The enrichment of true [synthetic lethal](#) partners among detect genes makes [SLIPT](#) valuable for triage of candidate [synthetic lethal](#) partners for further validation and for pathway analysis.

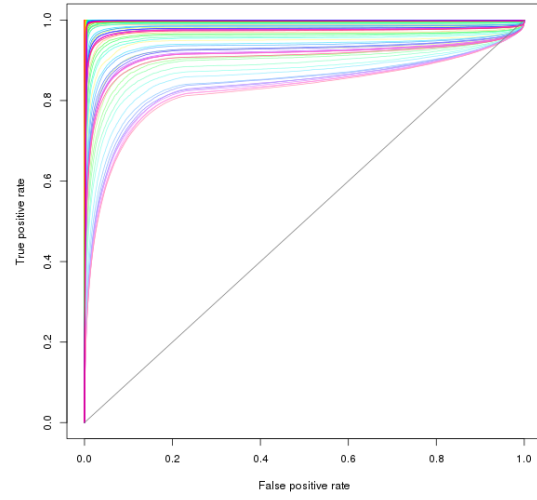
The performance of simulation of [synthetic lethality](#) within biological pathways was markedly higher in the context of a larger dataset of thousands of genes. As shown in a direct comparison with the [graph](#) structures alone (in Figure 6.20c), performance was consistently higher across parameters in pathways of biological complexity from the Reactome database (Croft *et al.*, 2014) such as [PI3K](#) cascade and the $G_{\alpha i}$ signalling pathway (shown in Figures K.18 and 6.20d).

These biologically complex [graph](#) structures, based on the Reactome pathways, assumed activating relationships to test [synthetic lethal](#) detection with [SLIPT](#) in the context of complex correlation structures. Inhibiting relationships were not distinguished in the Reactome database (Croft *et al.*, 2014). However, these investigations with pathway based [graph](#) structures indicate that the findings in constructed [graphs](#) (as used in Section 6.2) are relevant to [gene expression](#) data containing real correlated pathways. Furthermore, previous comparisons between simulations with inhibiting relationships indicated that the performance of [synthetic lethal](#) detection in an equivalent [graph](#) structure with inhibitory relationships will likely be higher.

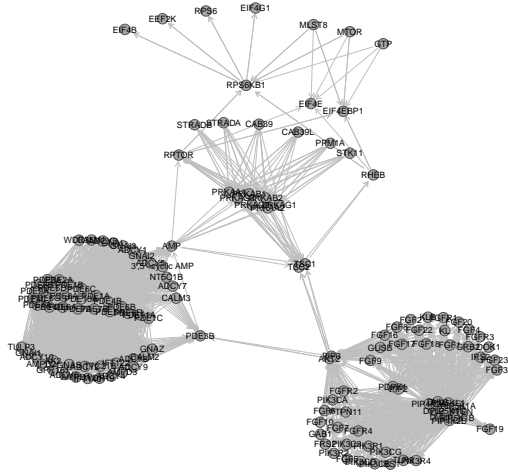
Non synthetic lethal genes, inversely correlated with the underlying [synthetic lethal](#) partners, were distinguishable by [SLIPT](#) with high specificity. [Synthetic lethal](#) genes were detectable with reasonable performance in large scale simulated [gene expression](#)



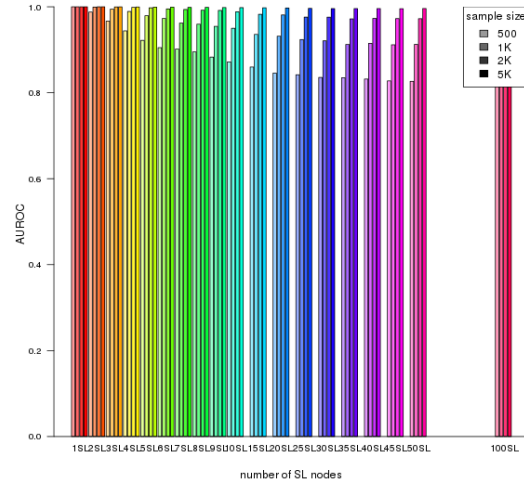
(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure 6.19: **Performance of simulations including the PI3K cascade.** Simulation of **synthetic lethality** was performed sampling from a multivariate normal distribution (without correlation structure apart from the Reactome **PI3K** cascade). Performance of **SLIPT** was high across parameters for detecting **synthetic lethality** in the **graph** structure within a larger dataset. The sensitivity decreases for a greater number of true positives to detect but the specificity remains high with a low false positive rate.

data and highly (positively) correlated genes in **pathway** structures. These findings serve as a conservative estimate for the performance of **SLIPT** to detect **synthetic lethal**

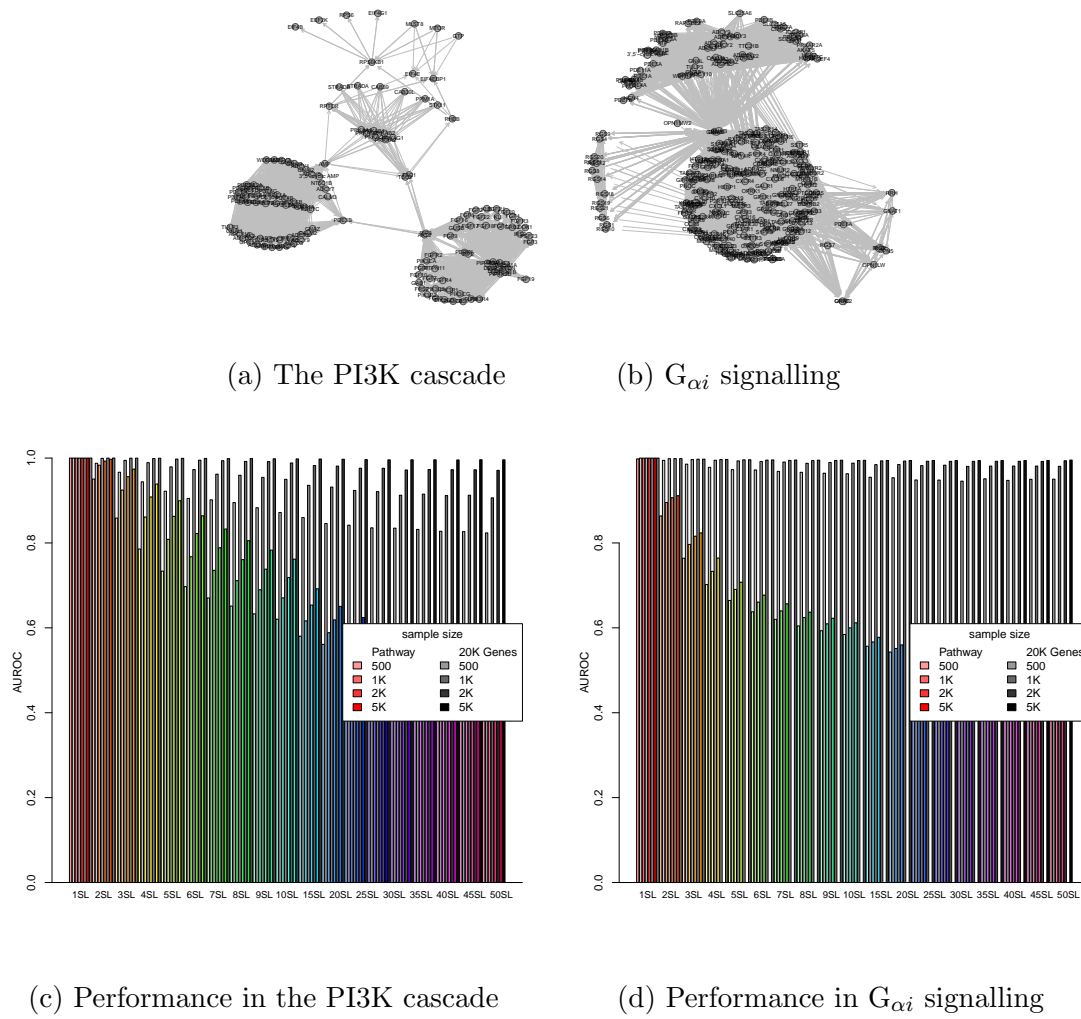


Figure 6.20: **Performance on pathways improves with more genes.** Simulations were performed in a [graph](#) structures for the [PI3K](#) cascade and $G_{\alpha i}$ signalling pathways structures to detect [synthetic lethal](#) partners within them. As for constructed graphs, performance of detection in a dataset containing only the [graph](#) structure (in colour) was as much lower than testing the [graph](#) structure within a larger dataset of non synthetic lethal genes (without correlations) for both [graphs](#) of biological complexity.

genes within a [synthetic lethal](#) biological pathway in empirical data. [Synthetic lethal](#) genes are distinguishable from correlated genes, to varying extents, in simulations. False positives are also more likely to occur within the same (synthetic lethal) pathways. Therefore [SLIPT](#) is both effective at triage of [synthetic lethal](#) candidates within a biological pathway and at identifying [synthetic lethal](#) pathways in high dimensional [gene expression](#) data.

6.4 Discussion

6.4.1 Simulation Procedure

Simulations have been performed to assess the performance of the [SLIPT](#) methodology (as described in Section 3.1 and with modifications) for detecting known underlying [synthetic lethal](#) partners of a query gene. These simulations support the findings in empirical data (in Chapters 4 and 5) by addressing whether the methodology used to generate them is accurate or has desirable statistical performance in controlled simulated conditions. These investigations include adjusting parameters such as the numbers of [synthetic lethal](#) genes which were known in empirical data to assess the performance of the [SLIPT](#) methodology across simulation parameters and characterise the datasets for which [SLIPT](#) performs well. Simulation and statistical modelling also enabled comparison of the [SLIPT](#) methodology to other statistical approaches to [synthetic lethal](#) detection in [expression](#) data.

These simulations are based on a statistical model of [synthetic lethality](#) (as described in Section 3.2.1) which was designed stringently to ensure that if [synthetic lethality](#) is detectable in the simulated datasets it would also be detectable by the same methodology in empirical [expression](#) data. The model of [synthetic lethality](#) made conservative assumptions such as the low threshold of [expression](#) for gene function or the inclusion of cryptic higher-order [synthetic lethality](#) (when testing pairwise). These assumptions decrease the likelihood that [synthetic lethal](#) signatures would be detectable in [expression](#) data. Thus it is reassuring that [synthetic lethality](#) is still detectable in under many simulation parameters as the performance of [SLIPT](#) would be expected to be higher were these assumptions to be violated in empirical data.

The simulation procedure (as described in Section 3.2.2) is designed as a computational pipeline with arguments passes to scripts. The [SLIPT](#) methodology and simulation of [expression](#) from [graph](#) structures were both used as R ([R Core Team, 2016](#)) software packages developed and released for this project (as described in Section 3.5). This design ensures that the simulations can be robustly applied across parameters with consistency between simulations apart from the differences discussed. The simulation procedure is also flexible to simulating other datasets, including [synthetic lethal](#) relationships and pathway correlation structures, should these be relevant to future investigations or [bioinformatics](#) tool development. The computational pipeline is also compatible with parallel computing and made use of [High Performance Computing \(HPC\)](#) infrastructure provided by [the New Zealand eScience Infrastructure \(NeSI\)](#) us-

ing the [Simple Linux Utility for Resource Management \(Slurm\)](#) submission system (as described in Section 2.5.3). This parallel computing pipeline enabled extensive investigations into [synthetic lethality](#) in simulated data, including approximately 2 million cpu-hours on [NeSI](#).

6.4.2 Comparing Methods with Simulated Data

Attempts were made to implement alternative [synthetic lethal](#) detection approaches such as linear regression and the [BiSEp](#) R package (discussed in Section 6.1). However, those tested were ineffective at detecting [synthetic lethality](#) in multivariate normal simulated data in comparison to [SLIPT](#). While some of the published [synthetic lethal](#) detection methods ([Jerby-Arnon *et al.*, 2014](#); [Lu *et al.*, 2015](#)) did not provide reproducible software releases for direct comparison, some of the central assumptions used in their design were tested by the statistical methods considered for [synthetic lethal](#) detection in [expression](#) data.

Another consideration is that [BiSEp](#) takes considerably more time to compute predictions than [SLIPT](#) or χ^2 which limited the number of simulations that were feasible and made it difficult to apply across parameters in the simulation pipeline (even when using supercomputing infrastructure as discussed in Section 2.5.3). The computationally intensive nature of the [BiSEp](#) procedure does not appear to be the issue for detecting [synthetic lethal](#) genes in [TCGA](#) data or simulations, although it has made more extensive simulations challenging. Rather, [BiSEp](#) is not suitable in either case since the [TCGA](#) data is normalised with [voom](#) ([Ritchie *et al.*, 2015](#)) and simulated data is generated by sampling from a multivariate normal distribution. In either case, even subtle bimodal signatures in [expression](#) data were not consistently detectable or sufficient to detect [synthetic lethality](#).

The [BiSEp](#) methodology may perform better on other data types but it cannot be directly compared with the results for [SLIPT](#) throughout this thesis which have used normalised or (multivariate) normally distributed data. Since it requires bimodal distributions, [BiSEp](#) is not suitable for stringently normalised [expression](#) data nor would it be expected to perform on (ranked) pathway [metagenes](#). Thus [SLIPT](#) represents a distinct approach more suitable for these data types whereas [BiSEp](#) may be applicable to other applications in which bimodal distributions are more frequent.

This investigation also demonstrates that implementing scientific software from other research groups is not a trivial exercise, even when released as an open-source R package. Therefore, the above results were used to evaluate [SLIPT](#) and compare it to

other statistical rationales. An comprehensive comparison to contemporary [synthetic lethal](#) detection approaches (and those released in the future) or further benchmarking is left to an impartial researcher to evaluate. The above findings show that the [SLIPT](#) approach is able to detect [synthetic lethal](#) genes in simulated data with comparable or better performance than a range of distinct statistical techniques and was appropriate for use throughout this thesis.

6.4.3 Design and Performance of SLIPT

The simulation procedure using sampling from a multivariate normal distribution was used throughout the majority of the simulation investigations in this thesis. This approach has the advantages of emulating the continuous normalised [expression](#) data used for [gene expression](#) analysis and enabled the simulation of correlation structures (as discussed in Section 3.3). These simulations scaled to datasets of comparable scale to those used in [gene expression](#) analysis with thousands of genes. The [SLIPT](#) methodology was shown to perform robustly across large numbers of genes and simple correlation structures. This includes high specificity against genes positively correlated with the query gene for which the directional [SLIPT](#) methodology more suited to distinguishing [synthetic lethal](#) genes from than the χ^2 test without directional criteria on the number of samples observed.

These findings were expanded upon in this chapter. Specifically, different quantiles were compared for [SLIPT](#) and the χ^2 test. These approaches using threshold based discrete gene function were compared to the Pearson correlation without loss of the continuous [expression](#) data. The 3-quantiles for [SLIPT](#) (as described in Section 3.1) were optimal for both [SLIPT](#) and the χ^2 alone. In addition to being optimal for estimating the significance of [synthetic lethal](#) interactions, these quantiles were also optimal for the directional criteria of [SLIPT](#) since this method outperformed the χ^2 test and was the most different at the 3-quantile. As previously, noted this difference was more pronounced with positively correlated genes (with the query gene) for which the specificity of [SLIPT](#) improves and was replicated in large datasets with thousands of genes as occur in human [expression](#) data. These results were not simply due to sufficient samples for significant p-values since the performance as determined by [AUROC](#) analysis is independent from significance thresholds. This indicates that the [SLIPT](#) methodology (as it has been used in Chapters 4 and 5) is optimal and the parameters used to design it were appropriate.

Both discrete functional approaches (SLIPT and χ^2) were able to outperform negative correlation which supports their use. In particular, this result addresses the concern that arbitrary thresholds of low and high gene function (as used by SLIPT) lose useful data by compressing the spectrum of gene expression into categorical data. However, this does not impede the performance of SLIPT and can reduce statistical if the quantiles used are optimal. The poorer performance of correlation-based detection of synthetic lethality also indicates affirms the concept of gene function for synthetic lethality being qualitative, that is expression must be sufficient for cell viability and higher expression is not necessarily higher function (as this is not the case for all genes). Furthermore, the finding that negative correlation outperforms positive correlation is also consistent with co-expression being a poor predictor of synthetic lethality compared to other approaches (Jerby-Arnon *et al.*, 2014), supporting the claims of Lu *et al.* (2015).

Compared with SLIPT, neither correlation approaches nor bimodality signatures were suitable for detecting synthetic lethality in expression data. The correlation-based approaches make assumptions about the relationship between gene expression and function which do not necessarily hold for all genes. Similarly, the bimodal approach is not appropriate for normalised data since deviations from a normal distribution have already been used for ensuring data quality, as is common practice for RNA-Seq data. Other approaches were continuous data such as fitting linear models are likely to be prone to similar issues and not perform as well as SLIPT. However, it is possible that these may be improved with conditioning on known synthetic lethal partners with multivariate regression or Bayesian priors. Similarly, synthetic lethal detection could be performed by iteratively conditioning upon the strong candidate from previous analysis. These approaches may be able to better circumvent the issues of high-order synthetic lethality and multiple testing.

Nevertheless, the above findings are sufficient to assess the performance of SLIPT and present an effective straightforward approach to synthetic lethal detection in gene expression data. Further development of linear models, Bayesian inference approaches, or comparison to existing synthetic lethal approaches (e.g., machine learning) remain as future directions. Developing and testing more sophisticated statistical approach to synthetic lethal detection may benefit from the concepts discussed with regard to the relatively simple SLIPT methodology. Similarly, further comparisons and benchmarking of SLIPT against other computational approaches to synthetic lethal detection in gene expression data is more suitable for an independent researcher and the `slipt` R

package has been released (as described in Section 3.5) for this purpose, in addition to further application in research.

6.4.4 Simulations from Graph Structures

The simple correlation structures (as used in Section 3.3) were expanded upon to simulate correlated genes based on [graph](#) structures using the multivariate normal simulation procedure on correlation structures generated from [graph](#) structures (as described in Section 6.2). These simulations enable further investigations into the performance of [SLIPT](#) in the context of more complex correlation structures. The simulation of [expression](#) from [network](#) structures is widely applicable to simulating pathway [expression](#) data and as such the [graphsim](#) R package has been released (as described in Section 3.5).

These investigations show that [SLIPT](#) performs robustly across datasets with different correlation structures, including those derived from [graphs](#) with the complexity of biological pathways. The [SLIPT](#) methodology was able to detect [synthetic lethal](#) genes within [synthetic lethal](#) pathways across many [graph](#) structures. This methodology performed particularly well with [synthetic lethal](#) pathways in the context of a larger dataset with a high specificity which supports [SLIPT](#) as a stringent approach to [synthetic lethal](#) detection in highly dimensional [gene expression](#) data. Together these results support the use of [SLIPT](#) in biological gene expression data since it is able to detect [synthetic lethal](#) genes in highly complex correlation structures.

Similarly, the inclusion of inhibitory relationships in [graph](#) structures was shown to increase the performance in simple networks supporting [SLIPT](#) being applicable to biological data in which these relationships are common. While these results were not replicated in more complex inhibitory [graph](#) structures, this is likely an artifact of the simulation procedure (which randomly selects [synthetic lethal](#) genes) generating biologically implausible combinations of [synthetic lethal](#) genes which are difficult to detect. When the test statistics in simulations with a [synthetic lethal](#) gene were examined in more detail, the test statistics of the [synthetic lethal](#) gene were consistently higher and distinguishable from nearby genes in the [graph](#) structure. In contrast to previous concerns with inhibiting relationships, these differences were more pronounced with genes which had inhibitory relationships with [synthetic lethal](#) genes. While distinguishable from nearby genes in a [pathway](#) structure, the genes correlated with [synthetic lethal](#) still had higher test statistics than more distant genes (similar to observations with correlated genes in Section 3.3).

In addition to being able to detect [synthetic lethal](#) genes in a pathway, the proximal genes in a pathway are most likely to be false positives and therefore [SLIPT](#) is also able to detect [synthetic lethal](#) pathways. Therefore [SLIPT](#) identifies genes which are likely to be constituent of a [synthetic lethal](#) pathway and is more likely to rank underlying [synthetic lethal](#) genes with greater significance. Together these findings support the use of [SLIPT](#) throughout this thesis, further application of [SLIPT](#), and further development of such strategies for [synthetic lethal](#) detection. Similarly, the simulation procedures developed and demonstrated for examining [synthetic lethal](#) detection in [expression](#) data using [graph](#) structures is amenable to further development and investigations into [pathway](#) structure in [expression](#) data such as predicting biological pathways from [expression](#) data or the impact of pathways on differential [expression](#) analyses.

6.5 Summary

A statistical model and simulation procedure has been developed to test the performance of the [SLIPT](#) methodology in controlled conditions, using multivariate normal distributions. This simulation procedure has been developed into a computational pipeline which was able to test the statistical performance (using stringent assumptions) of [SLIPT](#) across many parameters and compare it to alternative [synthetic lethal](#) detection strategies. The [SLIPT](#) methodology performs well at detecting small numbers of [synthetic lethal](#) genes in simple systems. It does not perform as well in more complex systems but neither do alternative strategies. The [SLIPT](#) methodology performs well compared to Pearson correlation and similar methods based on the χ^2 test. Thus [SLIPT](#) is an effective detection method for [synthetic lethal](#) relationships in [expression](#) data despite its relatively simple design.

Simulations of more complex datasets, including large numbers of genes, complex correlation structure derived from [graph](#) structures, and correlations with the query gene. [SLIPT](#) performs robustly across these, including correlation structures based on complex biological pathways. The performance of [SLIPT](#) improves in larger datasets, datasets with positive correlations with the query genes, and some [graph](#) structures which include inhibiting relationships, namely those datasets more representative of [gene expression](#) in biological data. [SLIPT](#) was both capable of recurrently detecting genes within a [synthetic lethal](#) pathway and distinguishing [synthetic lethal](#) genes from correlated with them, even in highly complex correlation structures. Therefore [SLIPT](#) is a stringent [synthetic lethal](#) detection strategy and is applicable to [gene expression](#)

as previously demonstrated for the partners of *CDH1* in breast and stomach cancer in this thesis.

Bibliography

- Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C.D., Annala, M., Aprikian, A., Armenia, J., Arora, A., *et al.* (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**(4): 1011–1025.
- Adler, D. (2005) *vioplot: Violin plot*. R package version 0.2.
- Akbani, R., Akdemir, K.C., Aksoy, B.A., Albert, M., Ally, A., Amin, S.B., Arachchi, H., Arora, A., Auman, J.T., Ayala, B., *et al.* (2015) Genomic Classification of Cutaneous Melanoma. *Cell*, **161**(7): 1681–1696.
- Akobeng, A.K. (2007) Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatrica*, **96**(5): 644–647.
- American Cancer Society (2017) Genetics and cancer. <https://www.cancer.org/cancer/cancer-causes/genetics.html>. Accessed: 22/03/2017.
- Anjomshoaa, A., Lin, Y.H., Black, M.A., McCall, J.L., Humar, B., Song, S., Fukuzawa, R., Yoon, H.S., Holzmann, B., Friederichs, J., *et al.* (2008) Reduced expression of a gene proliferation signature is associated with enhanced malignancy in colon cancer. *Br J Cancer*, **99**(6): 966–973.
- Araki, H., Knapp, C., Tsai, P., and Print, C. (2012) GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**: 76–82.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1): 25–29.
- Ashworth, A. (2008) A synthetic lethal therapeutic approach: poly(adp) ribose polymerase inhibitors for the treatment of cancers deficient in dna double-strand break repair. *J Clin Oncol*, **26**(22): 3785–90.

- Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011) Genetic interactions in cancer progression and treatment. *Cell*, **145**(1): 30–38.
- Audeh, M.W., Carmichael, J., Penson, R.T., Friedlander, M., Powell, B., Bell-McGuinn, K.M., Scott, C., Weitzel, J.N., Oaknin, A., Loman, N., *et al.* (2010) Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 245–51.
- Babyak, M.A. (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*, **66**(3): 411–21.
- Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, **91**(2): 355–358.
- Barabási, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439): 509–12.
- Barabási, A.L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**(1): 56–68.
- Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell’s functional organization. *Nat Rev Genet*, **5**(2): 101–13.
- Barrat, A. and Weigt, M. (2000) On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, **13**(3): 547–560.
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391): 603–607.
- Barry, W.T. (2016) *safe: Significance Analysis of Function and Expression*. R package version 3.14.0.

- Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B., and Boone, C. (2010a) Synthetic genetic array (sga) analysis in *saccharomyces cerevisiae* and *schizosaccharomyces pombe*. *Methods Enzymol*, **470**: 145–79.
- Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., *et al.* (2010b) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Meth*, **7**(12): 1017–1024.
- Bass, A.J., Thorsson, V., Shmulevich, I., Reynolds, S.M., Miller, M., Bernard, B., Hinoue, T., Laird, P.W., Curtis, C., Shen, H., *et al.* (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**(7517): 202–209.
- Bates, D. and Maechler, M. (2016) *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-7.1.
- Bateson, W. and Mendel, G. (1909) *Mendel's principles of heredity*, by W. Bateson. University Press, Cambridge [Eng.].
- Becker, K.F., Atkinson, M.J., Reich, U., Becker, I., Nekarda, H., Siewert, J.R., and Hfler, H. (1994) E-cadherin gene mutations provide clues to diffuse type gastric carcinomas. *Cancer Research*, **54**(14): 3845–3852.
- Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353): 609–615.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**(1): 289–300.
- Berx, G., Cleton-Jansen, A.M., Nollet, F., de Leeuw, W.J., van de Vijver, M., Cornelisse, C., and van Roy, F. (1995) E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. *EMBO J*, **14**(24): 6107–15.
- Berx, G., Cleton-Jansen, A.M., Strumane, K., de Leeuw, W.J., Nollet, F., van Roy, F., and Cornelisse, C. (1996) E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain. *Oncogene*, **13**(9): 1919–25.

- Berx, G. and van Roy, F. (2009) Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harb Perspect Biol*, **1**: a003129.
- Bitler, B.G., Aird, K.M., Garipov, A., Li, H., Amatangelo, M., Kossenkov, A.V., Schultz, D.C., Liu, Q., Shih Ie, M., Conejo-Garcia, J.R., *et al.* (2015) Synthetic lethality by targeting ezh2 methyltransferase activity in arid1a-mutated cancers. *Nat Med*, **21**(3): 231–8.
- Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Burgess, S., Buza, T., Gresham, C., *et al.* (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res*, **43**(Database issue): D1049–1056.
- Boone, C., Bussey, H., and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, **8**(6): 437–49.
- Borgatti, S.P. (2005) Centrality and network flow. *Social Networks*, **27**(1): 55 – 71.
- Boucher, B. and Jenna, S. (2013) Genetic interaction networks: better understand to better predict. *Front Genet*, **4**: 290.
- Bozovic-Spasojevic, I., Azambuja, E., McCaskill-Stevens, W., Dinh, P., and Cardoso, F. (2012) Chemoprevention for breast cancer. *Cancer treatment reviews*, **38**(5): 329–339.
- Breiman, L. (2001) Random forests. *Machine Learning*, **45**(1): 5–32.
- Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, **30**(1): 107 – 117.
- Brouxhon, S.M., Kyrkanides, S., Teng, X., Athar, M., Ghazizadeh, S., Simon, M., O’Banion, M.K., and Ma, L. (2014) Soluble E-cadherin: a critical oncogene modulating receptor tyrosine kinases, MAPK and PI3K/Akt/mTOR signaling. *Oncogene*, **33**(2): 225–235.
- Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009) Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*, **10**(6): 2763–2788.
- Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., and Helleday, T. (2005) Specific killing of *BRCA2*-deficient tumours with inhibitors of polyadpribose polymerase. *Nature*, **434**(7035): 913–7.

- Bussey, H., Andrews, B., and Boone, C. (2006) From worm genetic networks to complex human diseases. *Nat Genet*, **38**(8): 862–3.
- Butland, G., Babu, M., Diaz-Mejia, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O., *et al.* (2008) esga: E. coli synthetic genetic array analysis. *Nat Methods*, **5**(9): 789–95.
- cBioPortal for Cancer Genomics (cBioPortal) (2017) cBioPortal for Cancer Genomics. <http://www.cbioportal.org/>. Accessed: 26/03/2017.
- Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, **39**(Database issue): D685–690.
- Chen, A., Beetham, H., Black, M.A., Priya, R., Telford, B.J., Guest, J., Wiggins, G.A.R., Godwin, T.D., Yap, A.S., and Guilford, P.J. (2014) E-cadherin loss alters cytoskeletal organization and adhesion in non-malignant breast cells but is insufficient to induce an epithelial-mesenchymal transition. *BMC Cancer*, **14**(1): 552.
- Chen, S. and Parmigiani, G. (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol*, **25**(11): 1329–1333.
- Chipman, K. and Singh, A. (2009) Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, **10**(1): 17.
- Christofori, G. and Semb, H. (1999) The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends in Biochemical Sciences*, **24**(2): 73 – 76.
- Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., *et al.* (2015) Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, **163**(2): 506–519.
- Clark, M.J. (2004) Endogenous Regulator of G Protein Signaling Proteins Suppress G o-Dependent μ -Opioid Agonist-Mediated Adenylyl Cyclase Supersensitization. *Journal of Pharmacology and Experimental Therapeutics*, **310**(1): 215–222.
- Collingridge, D.S. (2013) A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, **7**(1): 81–97.

- Collins, F.S. and Barker, A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am*, **296**(3): 50–57.
- Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer, D., Cope, L., Creighton, C., Danilova, L., *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**(7511): 543–550.
- Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., *et al.* (2010) The genetic landscape of a cell. *Science*, **327**(5964): 425–31.
- Costanzo, M., Baryshnikova, A., Myers, C.L., Andrews, B., and Boone, C. (2011) Charting the genetic interaction map of a cell. *Curr Opin Biotechnol*, **22**(1): 66–74.
- Courtney, K.D., Corcoran, R.B., and Engelman, J.A. (2010) The PI3K pathway as drug target in human cancer. *J Clin Oncol*, **28**(6): 1075–1083.
- Creighton, C.J., Morgan, M., Gunaratne, P.H., Wheeler, D.A., Gibbs, R.A., Robertson, A., Chu, A., Beroukhim, R., Cibulskis, K., Signoretti, S., *et al.* (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**(7456): 43–49.
- Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.* (2014) The Reactome pathway knowledge-base. *Nucleic Acids Res*, **42**(database issue): D472D477.
- Crunkhorn, S. (2014) Cancer: Predicting synthetic lethal interactions. *Nat Rev Drug Discov*, **13**(11): 812.
- Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Systems**: 1695.
- Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res*, **5**(10): 2929–2943.
- Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y., *et al.* (2005) The synthetic genetic interaction spectrum of essential genes. *Nat Genet*, **37**(10): 1147–1152.

- De Leeuw, W.J., Berx, G., Vos, C.B., Peterse, J.L., Van de Vijver, M.J., Litvinov, S., Van Roy, F., Cornelisse, C.J., and Cleton-Jansen, A.M. (1997) Simultaneous loss of E-cadherin and catenins in invasive lobular breast cancer and lobular carcinoma in situ. *J Pathol*, **183**(4): 404–11.
- De Santis, G., Miotti, S., Mazzi, M., Canevari, S., and Tomassetti, A. (2009) E-cadherin directly contributes to PI3K/AKT activation by engaging the PI3K-p85 regulatory subunit to adherens junctions of ovarian carcinoma cells. *Oncogene*, **28**(9): 1206–1217.
- Demir, E., Babur, O., Rodchenkov, I., Aksoy, B.A., Fukuda, K.I., Gross, B., Sumer, O.S., Bader, G.D., and Sander, C. (2013) Using biological pathway data with Paxtools. *PLoS Comput Biol*, **9**(9): e1003194.
- Deshpande, R., Asiedu, M.K., Klebig, M., Sutor, S., Kuzmin, E., Nelson, J., Piotrowski, J., Shin, S.H., Yoshida, M., Costanzo, M., *et al.* (2013) A comparative genomic approach for identifying synthetic lethal interactions in human cancer. *Cancer Res*, **73**(20): 6128–36.
- Dickson, D. (1999) Wellcome funds cancer database. *Nature*, **401**(6755): 729.
- Dijkstra, E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**(1): 269–271.
- Dixon, S.J., Andrews, B.J., and Boone, C. (2009) Exploring the conservation of synthetic lethal genetic interaction networks. *Commun Integr Biol*, **2**(2): 78–81.
- Dixon, S.J., Fedyshyn, Y., Koh, J.L., Prasad, T.S., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.L., *et al.* (2008) Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*, **105**(43): 16653–8.
- Dong, L.L., Liu, L., Ma, C.H., Li, J.S., Du, C., Xu, S., Han, L.H., Li, L., and Wang, X.W. (2012) E-cadherin promotes proliferation of human ovarian cancer cells in vitro via activating MEK/ERK pathway. *Acta Pharmacol Sin*, **33**(6): 817–822.
- Dorsam, R.T. and Gutkind, J.S. (2007) G-protein-coupled receptors and cancer. *Nat Rev Cancer*, **7**(2): 79–94.
- Erdős, P. and Rényi, A. (1959) On random graphs I. *Publ Math Debrecen*, **6**: 290–297.

- Erdős, P. and Rényi, A. (1960) On the evolution of random graphs. In *Publ. Math. Inst. Hung. Acad. Sci*, volume 5, 17–61.
- Eroles, P., Bosch, A., Perez-Fidalgo, J.A., and Lluch, A. (2012) Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer Treat Rev*, **38**(6): 698–707.
- Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., *et al.* (2005) Targeting the dna repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**(7035): 917–21.
- Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8): 861 – 874. {ROC} Analysis in Pattern Recognition.
- Fece de la Cruz, F., Gapp, B.V., and Nijman, S.M. (2015) Synthetic lethal vulnerabilities of cancer. *Annu Rev Pharmacol Toxicol*, **55**: 513–531.
- Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**(5): E359–386.
- Fisher, R.A. (1919) Xv.the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52**(02): 399–433.
- Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O’Connor, M.J., *et al.* (2009) Inhibition of poly(adp-ribose) polymerase in tumors from BRCA mutation carriers. *N Engl J Med*, **361**(2): 123–34.
- Fong, P.C., Yap, T.A., Boss, D.S., Carden, C.P., Mergui-Roelvink, M., Gourley, C., De Greve, J., Lubinski, J., Shanley, S., Messiou, C., *et al.* (2010) Poly(adp)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. *J Clin Oncol*, **28**(15): 2512–9.
- Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., *et al.* (2015) COSMIC: exploring the world’s

- knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**(Database issue): D805–811.
- Fraser, A. (2004) Towards full employment: using RNAi to find roles for the redundant. *Oncogene*, **23**(51): 8346–52.
- Fromental-Ramain, C., Warot, X., Lakkaraju, S., Favier, B., Haack, H., Birling, C., Dierich, A., Doll e, P., and Chambon, P. (1996) Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning. *Development*, **122**(2): 461–472.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004) A census of human cancer genes. *Nat Rev Cancer*, **4**(3): 177–183.
- Futreal, P.A., Kasprzyk, A., Birney, E., Mullikin, J.C., Wooster, R., and Stratton, M.R. (2001) Cancer and genomics. *Nature*, **409**(6822): 850–852.
- Gao, B. and Roux, P.P. (2015) Translational control by oncogenic signaling pathways. *Biochimica et Biophysica Acta*, **1849**(7): 753–65.
- Gatza, M.L., Kung, H.N., Blackwell, K.L., Dewhirst, M.W., Marks, J.R., and Chi, J.T. (2011) Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes. *Breast Cancer Res*, **13**(3): R62.
- Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A., *et al.* (2010) A pathway-based classification of human breast cancer. *Proc Natl Acad Sci USA*, **107**(15): 6994–6999.
- Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C., and Perou, C.M. (2014) An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet*, **46**(10): 1051–1059.
- Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10): R80.
- Genz, A. and Bretz, F. (2009) Computation of multivariate normal and t probabilities. In *Lecture Notes in Statistics*, volume 195. Springer-Verlag, Heidelberg.

- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2016) *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5. URL.
- Glaire, M.A., Brown, M., Church, D.N., and Tomlinson, I. (2017) Cancer predisposition syndromes: lessons for truly precision medicine. *J Pathol*, **241**(2): 226–235.
- Globus (Globus) (2017) Research data management simplified. <https://www.globus.org/>. Accessed: 25/03/2017.
- Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**(6): 333–351.
- Grady, W.M., Willis, J., Guilford, P.J., Dunbier, A.K., Toro, T.T., Lynch, H., Wiesner, G., Ferguson, K., Eng, C., Park, J.G., *et al.* (2000) Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat Genet*, **26**(1): 16–17.
- Graziano, F., Humar, B., and Guilford, P. (2003) The role of the E-cadherin gene (*CDH1*) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice. *Annals of Oncology*, **14**(12): 1705–1713.
- Guaragnella, N., Palermo, V., Galli, A., Moro, L., Mazzoni, C., and Giannattasio, S. (2014) The expanding role of yeast in cancer research and diagnosis: insights into the function of the oncosuppressors p53 and BRCA1/2. *FEMS Yeast Res*, **14**(1): 2–16.
- Güell, O., Sagus, F., and Serrano, M. (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. *PLoS Comput Biol*, **10**(5): e1003637.
- Guilford, P. (1999) E-cadherin downregulation in cancer: fuel on the fire? *Molecular Medicine Today*, **5**(4): 172 – 177.
- Guilford, P., Hopkins, J., Harraway, J., McLeod, M., McLeod, N., Harawira, P., Taite, H., Scoular, R., Miller, A., and Reeve, A.E. (1998) E-cadherin germline mutations in familial gastric cancer. *Nature*, **392**(6674): 402–5.
- Guilford, P., Humar, B., and Blair, V. (2010) Hereditary diffuse gastric cancer: translation of *CDH1* germline mutations into clinical practice. *Gastric Cancer*, **13**(1): 1–10.

- Guilford, P.J., Hopkins, J.B., Grady, W.M., Markowitz, S.D., Willis, J., Lynch, H., Rajput, A., Wiesner, G.L., Lindor, N.M., Burgart, L.J., *et al.* (1999) E-cadherin germline mutations define an inherited cancer syndrome dominated by diffuse gastric cancer. *Hum Mutat*, **14**(3): 249–55.
- Guo, J., Liu, H., and Zheng, J. (2016) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res*, **44**(D1): D1011–1017.
- Hajian-Tilaki, K. (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, **4**(2): 627–635.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009) The weka data mining software: an update. *SIGKDD Explor Newsl*, **11**(1): 10–18.
- Hammerman, P.S., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., *et al.* (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**(7417): 519–525.
- Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**(1): 57–70.
- Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**(5): 646–674.
- Hanna, S. (2003) Cancer incidence in new zealand (2003-2007). In D. Forman, D. Bray F Brewster, C. Gombe Mbalawa, B. Kohler, M. Piñeros, E. Steliarova-Foucher, R. Swaminathan, and J. Ferlay (editors), *Cancer Incidence in Five Continents*, volume X, 902–907. International Agency for Research on Cancer, Lyon, France. Electronic version <http://ci5.iarc.fr> Accessed 22/03/2017.
- Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., Schrader, K.A., Schaeffer, D.F., Shumansky, K., Zogopoulos, G., *et al.* (2015) Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol*, **1**(1): 23–32.
- Heiskanen, M.A. and Aittokallio, T. (2012) Mining high-throughput screens for cancer drug targets-lessons from yeast chemical-genomic profiling and synthetic lethality.

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **2**(3): 263–272.

Hell, P. (1976) Graphs with given neighbourhoods i. problèmes combinatorics at theorie des graphes. *Proc Coil Int CNRS, Orsay*, **260**: 219–223.

Higgins, M.E., Claremont, M., Major, J.E., Sander, C., and Lash, A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res*, **35**(Database issue): D721–726.

Hillenmeyer, M.E. (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, **320**: 362–365.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V., *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**(4): 929–944.

Hoehndorf, R., Hardy, N.W., Osumi-Sutherland, D., Tweedie, S., Schofield, P.N., and Gkoutos, G.V. (2013) Systematic analysis of experimental phenotype data reveals gene functions. *PLoS ONE*, **8**(4): e60847.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**(2): 65–70.

Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, **4**(11): 682–690.

Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**: 96.

Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., *et al.* (2003) Gene expression predictors of breast cancer outcomes. *Lancet*, **361**: 1590–1596.

Hutchison, C.A., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science*, **351**(6280): aad6253.

- International HapMap 3 Consortium (HapMap) (2003) The International HapMap Project. *Nature*, **426**(6968): 789–796.
- Jeanes, A., Gottardi, C.J., and Yap, A.S. (2008) Cadherins and cancer: how does cadherin dysfunction promote tumor progression? *Oncogene*, **27**(55): 6920–6929.
- Jerby-Arnon, L., Pfetzer, N., Waldman, Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P., *et al.* (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, **158**(5): 1199–1209.
- Joachims, T. (1999) Making large-scale support vector machine learning practical. In S. Bernhard, I. Kopr, J.C.B. Christopher, and J.S. Alexander (editors), *Advances in kernel methods*, 169–184. MIT Press.
- Ju, Z., Liu, W., Roebuck, P.L., Siwak, D.R., Zhang, N., Lu, Y., Davies, M.A., Akbani, R., Weinstein, J.N., Mills, G.B., *et al.* (2015) Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics*, **31**(6): 912.
- Kaelin, Jr, W. (2005) The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer*, **5**(9): 689–98.
- Kaelin, Jr, W. (2009) Synthetic lethality: a framework for the development of wiser cancer therapeutics. *Genome Med*, **1**: 99.
- Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**(1): 7–15.
- Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**(6821): 685–690.
- Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotech*, **23**(5): 561–566.
- Kelly, S.T. (2013) *Statistical Predictions of Synthetic Lethal Interactions in Cancer*. Dissertation, University of Otago.
- Kelly, S.T., Single, A.B., Telford, B.J., Beetham, H.G., Godwin, T.D., Chen, A., Black, M.A., and Guilford, P.J. (unpublished) Towards HDGC chemoprevention:

- vulnerabilities in E-cadherin-negative cells identified by genome-wide interrogation of isogenic cell lines and whole tumors. Submitted to *Cancer Prev Res*.
- Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res*, **37**(Database issue): D767–772.
- Kim, N.G., Koh, E., Chen, X., and Gumbiner, B.M. (2011) E-cadherin mediates contact inhibition of proliferation through Hippo signaling-pathway components. *Proc Natl Acad Sci USA*, **108**(29): 11930–11935.
- Kockel, L., Zeitlinger, J., Staszewski, L.M., Mlodzik, M., and Bohmann, D. (1997) Jun in drosophila development: redundant and nonredundant functions and regulation by two mapk signal transduction pathways. *Genes & Development*, **11**(13): 1748–1758.
- Kozlov, K.N., Gursky, V.V., Kulakovskiy, I.V., and Samsonova, M.G. (2015) Sequence-based model of gap gene regulation network. *BMC Genomics*, **15**(Suppl 12): S6.
- Kranthi, S., Rao, S., and Manimaran, P. (2013) Identification of synthetic lethal pairs in biological systems through network information centrality. *Mol BioSyst*, **9**(8): 2163–2167.
- Kroepil, F., Fluegen, G., Totikov, Z., Baldus, S.E., Vay, C., Schauer, M., Topp, S.A., Esch, J.S., Knoefel, W.T., and Stoecklein, N.H. (2012) Down-regulation of CDH1 is associated with expression of SNAIL in colorectal adenomas. *PLoS ONE*, **7**(9): e46665.
- Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**(7333): 187–197.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822): 860–921.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3): R25.

- Latora, V. and Marchiori, M. (2001) Efficient behavior of small-world networks. *Phys Rev Lett*, **87**: 198701.
- Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013) Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat Methods*, **10**(5): 427–31.
- Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, **15**(2): R29.
- Le Meur, N. and Gentleman, R. (2008) Modeling synthetic lethality. *Genome Biol*, **9**(9): R135.
- Le Meur, N., Jiang, Z., Liu, T., Mar, J., and Gentleman, R.C. (2014) Slgi: Synthetic lethal genetic interaction. r package version 1.26.0.
- Lee, A.Y., Perreault, R., Harel, S., Boulrier, E.L., Suderman, M., Hallett, M., and Jenna, S. (2010a) Searching for signaling balance through the identification of genetic interactors of the rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS ONE*, **5**(5): e10624.
- Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G., and Marcotte, E.M. (2010b) Predicting genetic modifier loci using functional gene networks. *Genome Research*, **20**(8): 1143–1153.
- Lee, I. and Marcotte, E.M. (2009) Effects of functional bias on supervised learning of a gene network model. *Methods Mol Biol*, **541**: 463–75.
- Lee, M.J., Ye, A.S., Gardino, A.K., Heijink, A.M., Sorger, P.K., MacBeath, G., and Yaffe, M.B. (2012) Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell*, **149**(4): 780–94.
- Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006) Systematic mapping of genetic interactions in *caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat Genet*, **38**(8): 896–903.
- Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4): 493–500.

- Li, X.J., Mishra, S.K., Wu, M., Zhang, F., and Zheng, J. (2014) Syn-lethality: An integrative knowledge base of synthetic lethality towards discovery of selective anti-cancer therapies. *Biomed Res Int*, **2014**: 196034.
- Linehan, W.M., Spellman, P.T., Ricketts, C.J., Creighton, C.J., Fei, S.S., Davis, C., Wheeler, D.A., Murray, B.A., Schmidt, L., Vocke, C.D., *et al.* (2016) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med*, **374**(2): 135–145.
- Lokody, I. (2014) Computational modelling: A computational crystal ball. *Nature Reviews Cancer*, **14**(10): 649–649.
- Lord, C.J., Tutt, A.N., and Ashworth, A. (2015) Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annu Rev Med*, **66**: 455–470.
- Lu, X., Kensche, P.R., Huynen, M.A., and Notebaart, R.A. (2013) Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nat Commun*, **4**: 2124.
- Lu, X., Megchelenbrink, W., Notebaart, R.A., and Huynen, M.A. (2015) Predicting human genetic interactions from cancer genome evolution. *PLoS One*, **10**(5): e0125795.
- Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., Wolf, M.K., Butler, J.S., Hinshaw, J.C., Garnier, P., Prestwich, G.D., Leonardson, A., *et al.* (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**(1): 121–137.
- Luo, J., Solimini, N.L., and Elledge, S.J. (2009) Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell*, **136**(5): 823–837.
- Machado, J., Olivera, C., Carvalh, R., Soares, P., Berx, G., Caldas, C., Sercuca, R., Carneiro, F., and Sorbrinho-Simoes, M. (2001) E-cadherin gene (*CDH1*) promoter methylation as the second hit in sporadic diffuse gastric carcinoma. *Oncogene*, **20**: 1525–1528.
- Markowetz, F. (2017) All biology is computational biology. *PLoS Biol*, **15**(3): e2002050.

- Masciari, S., Larsson, N., Senz, J., Boyd, N., Kaurah, P., Kandel, M.J., Harris, L.N., Pinheiro, H.C., Troussard, A., Miron, P., *et al.* (2007) Germline E-cadherin mutations in familial lobular breast cancer. *J Med Genet*, **44**(11): 726–31.
- Mattison, J., van der Weyden, L., Hubbard, T., and Adams, D.J. (2009) Cancer gene discovery in mouse and man. *Biochim Biophys Acta*, **1796**(2): 140–161.
- McLachlan, J., George, A., and Banerjee, S. (2016) The current status of parp inhibitors in ovarian cancer. *Tumori*, **102**(5): 433–440.
- McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., Mastrogianakis, G.M., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216): 1061–1068.
- Miles, D.W. (2001) Update on HER-2 as a target for cancer therapy: herceptin in the clinical setting. *Breast Cancer Res*, **3**(6): 380–384.
- Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., *et al.* (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407): 330–337.
- Nagalla, S., Chou, J.W., Willingham, M.C., Ruiz, J., Vaughn, J.P., Dubey, P., Lash, T.L., Hamilton-Dutoit, S.J., Bergh, J., Sotiriou, C., *et al.* (2013) Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol*, **14**(4): R34.
- Neeley, E.S., Kornblau, S.M., Coombes, K.R., and Baggerly, K.A. (2009) Variable slope normalization of reverse phase protein arrays. *Bioinformatics*, **25**(11): 1384.
- Novomestky, F. (2012) *matrixcalc: Collection of functions for matrix calculations*. R package version 1.0-3.
- Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997) Evolution of genetic redundancy. *Nature*, **388**(6638): 167–171.
- Oliveira, C., Senz, J., Kaurah, P., Pinheiro, H., Sanges, R., Haegert, A., Corso, G., Schouten, J., Fitzgerald, R., Vogelsang, H., *et al.* (2009) Germline *CDH1* deletions in hereditary diffuse gastric cancer families. *Human Molecular Genetics*, **18**(9): 1545–1555.

- Oliveira, C., Seruca, R., Hoogerbrugge, N., Ligtenberg, M., and Carneiro, F. (2013) Clinical utility gene card for: Hereditary diffuse gastric cancer (HDGC). *Eur J Hum Genet*, **21**(8).
- Pandey, G., Zhang, B., Chang, A.N., Myers, C.L., Zhu, J., Kumar, V., and Schadt, E.E. (2010) An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol*, **6**(9).
- Parker, J., Mullins, M., Cheung, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8): 1160–1167.
- Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., *et al.* (2016) Erratum: The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*, **7**: 11908.
- Perou, C.M., Sørlie, 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., *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**(6797): 747–752.
- Polyak, K. and Weinberg, R.A. (2009) Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer*, **9**(4): 265–73.
- R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.3.2.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7): e47.
- Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., *et al.* (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, **322**(5900): 405–10.
- Roychowdhury, S. and Chinnaiyan, A.M. (2016) Translating cancer genomes and transcriptomes for precision oncology. *CA Cancer J Clin*, **66**(1): 75–88.
- Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. *Nat Rev Genet*, **14**(2): 89–99.

- Ryan, C., Lord, C., and Ashworth, A. (2014) Daisy: Picking synthetic lethals from cancer genomes. *Cancer Cell*, **26**(3): 306–308.
- Schena, M. (1996) Genome analysis with gene expression microarrays. *Bioessays*, **18**(5): 427–431.
- Scheuer, L., Kauff, N., Robson, M., Kelly, B., Barakat, R., Satagopan, J., Ellis, N., Hensley, M., Boyd, J., Borgen, P., *et al.* (2002) Outcome of preventive surgery and screening for breast and ovarian cancer in BRCA mutation carriers. *J Clin Oncol*, **20**(5): 1260–1268.
- Semb, H. and Christofori, G. (1998) The tumor-suppressor function of E-cadherin. *Am J Hum Genet*, **63**(6): 1588–93.
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) Rocr: visualizing classifier performance in r. *Bioinformatics*, **21**(20): 7881.
- Slurm development team (Slurm) (2017) Slurm workload manager. <https://slurm.schedmd.com/>. Accessed: 25/03/2017.
- Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*, **98**(19): 10869–10874.
- Stajich, J.E. and Lapp, H. (2006) Open source tools and toolkits for bioinformatics: significance, and where are we? *Brief Bioinformatics*, **7**(3): 287–296.
- Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009) The cancer genome. *Nature*, **458**(7239): 719–724.
- Ström, C. and Helleday, T. (2012) Strategies for the use of poly(adenosine diphosphate ribose) polymerase (parp) inhibitors in cancer therapy. *Biomolecules*, **2**(4): 635–649.
- Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res*, **21**(12): 2213–2223.
- Telford, B.J., Chen, A., Beetham, H., Frick, J., Brew, T.P., Gould, C.M., Single, A., Godwin, T., Simpson, K.J., and Guilford, P. (2015) Synthetic lethal screens identify vulnerabilities in gpcr signalling and cytoskeletal organization in E-cadherin-deficient cells. *Mol Cancer Ther*, **14**(5): 1213–1223.

- The 1000 Genomes Project Consortium (1000 Genomes) (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319): 1061–1073.
- The Cancer Genome Atlas Research Network (TCGA) (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418): 61–70.
- The Cancer Genome Atlas Research Network (TCGA) (2017) The Cancer Genome Atlas Project. <https://cancergenome.nih.gov/>. Accessed: 26/03/2017.
- The Catalogue Of Somatic Mutations In Cancer (COSMIC) (2016) Cosmic: The catalogue of somatic mutations in cancer. <http://cancer.sanger.ac.uk/cosmic>. Release 79 (23/08/2016), Accessed: 05/02/2017.
- The Comprehensive R Archive Network (CRAN) (2017) Cran. <https://cran.r-project.org/>. Accessed: 24/03/2017.
- The ENCODE Project Consortium (ENCODE) (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**(5696): 636–640.
- The National Cancer Institute (NCI) (2015) The genetics of cancer. <https://www.cancer.gov/about-cancer/causes-prevention/genetics>. Published: 22/04/2015, Accessed: 22/03/2017.
- The New Zealand eScience Infrastructure (NeSI) (2017) NeSI. <https://www.nesi.org.nz/>. Accessed: 25/03/2017.
- Tierney, L., Rossini, A.J., Li, N., and Sevcikova, H. (2015) *snow: Simple Network of Workstations*. R package version 0.4-2.
- Tiong, K.L., Chang, K.C., Yeh, K.T., Liu, T.Y., Wu, J.H., Hsieh, P.H., Lin, S.H., Lai, W.Y., Hsu, Y.C., Chen, J.Y., *et al.* (2014) Csnk1e/ctnnb1 are synthetic lethal to tp53 in colorectal cancer and are markers for prognosis. *Neoplasia*, **16**(5): 441–50.
- Tischler, J., Lehner, B., and Fraser, A.G. (2008) Evolutionary plasticity of genetic interaction networks. *Nat Genet*, **40**(4): 390–391.
- Tomasetti, C. and Vogelstein, B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**(6217): 78–81.

- Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**(5550): 2364–8.
- Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**(5659): 808–13.
- Tran, B., Dancey, J.E., Kamel-Reid, S., McPherson, J.D., Bedard, P.L., Brown, A.M., Zhang, T., Shaw, P., Onetto, N., Stein, L., *et al.* (2012) Cancer genomics: technology, discovery, and translation. *J Clin Oncol*, **30**(6): 647–660.
- Travers, J. and Milgram, S. (1969) An experimental study of the small world problem. *Sociometry*, **32**(4): 425–443.
- Tunggal, J.A., Helfrich, I., Schmitz, A., Schwarz, H., Gunzel, D., Fromm, M., Kemler, R., Krieg, T., and Niessen, C.M. (2005) E-cadherin is essential for in vivo epidermal barrier function by regulating tight junctions. *EMBO J*, **24**(6): 1146–1156.
- 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., *et al.* (2010) Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and advanced breast cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 235–44.
- University of California, Santa Cruz (UCSC) (2012) Usc cancer browser. Accessed 29/03/2012.
- van der Post, R.S., Vogelaar, I.P., Carneiro, F., Guilford, P., Huntsman, D., Hoogerbrugge, N., Caldas, C., Schreiber, K.E., Hardwick, R.H., Ausems, M.G., *et al.* (2015) Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. *J Med Genet*, **52**(6): 361–374.
- van Steen, K. (2012) Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, **13**(1): 1–19.
- van Steen, M. (2010) *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen, VU Amsterdam.
- Vapnik, V.N. (1995) *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

- Vizeacoumar, F.J., Arnold, R., Vizeacoumar, F.S., Chandrashekhar, M., Buzina, A., Young, J.T., Kwan, J.H., Sayad, A., Mero, P., Lawo, S., *et al.* (2013) A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Mol Syst Biol*, **9**: 696.
- Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**(6127): 1546–1558.
- Vos, C.B., Cleton-Jansen, A.M., Berx, G., de Leeuw, W.J., ter Haar, N.T., van Roy, F., Cornelisse, C.J., Peterse, J.L., and van de Vijver, M.J. (1997) E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. *Br J Cancer*, **76**(9): 1131–3.
- Waldron, D. (2016) Cancer genomics: A multi-layer omics approach to cancer. *Nat Rev Genet*, **17**(8): 436–437.
- Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, **38**(18): e178.
- Wang, X. and Simon, R. (2013) Identification of potential synthetic lethal genes to p53 using a computational biology approach. *BMC Medical Genomics*, **6**(1): 30.
- Wappett, M. (2014) Bisep: Toolkit to identify candidate synthetic lethality. r package version 2.0.
- Wappett, M., Dulak, A., Yang, Z.R., Al-Watban, A., Bradford, J.R., and Dry, J.R. (2016) Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs. *BMC Genomics*, **17**: 65.
- Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., *et al.* (2015) *gplots: Various R Programming Tools for Plotting Data*. R package version 2.17.0.
- Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**(6684): 440–2.
- Weinstein, I.B. (2000) Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis. *Carcinogenesis*, **21**(5): 857–864.

- Weinstein, J.N., Akbani, R., Broom, B.M., Wang, W., Verhaak, R.G., McConkey, D., Lerner, S., Morgan, M., Creighton, C.J., Smith, C., *et al.* (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, **507**(7492): 315–322.
- Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Chang, K., *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, **45**(10): 1113–1120.
- Wickham, H. and Chang, W. (2016) *devtools: Tools to Make Developing R Packages Easier*. R package version 1.12.0.
- Wickham, H., Danenberg, P., and Eugster, M. (2017) *roxygen2: In-Line Documentation for R*. R package version 6.0.1.
- Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., *et al.* (2004) Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(44): 15682–15687.
- World Health Organization (WHO) (2017) Fact sheet: Cancer. <http://www.who.int/mediacentre/factsheets/fs297/en/>. Updated February 2017, Accessed: 22/03/2017.
- Wu, M., Li, X., Zhang, F., Li, X., Kwoh, C.K., and Zheng, J. (2014) In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. *Cancer Inform*, **13**(Suppl 3): 71–80.
- Yu, H. (2002) Rmpi: Parallel statistical computing in r. *R News*, **2**(2): 10–14.
- Zhang, F., Wu, M., Li, X.J., Li, X.L., Kwoh, C.K., and Zheng, J. (2015) Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *J Bioinform Comput Biol*, **13**(3): 1541002.
- Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., *et al.* (2011) International cancer genome consortium data portala one-stop shop for cancer genomics data. *Database: The Journal of Biological Databases and Curation*, **2011**: bar026.
- Zhong, W. and Sternberg, P.W. (2006) Genome-wide prediction of c. elegans genetic interactions. *Science*, **311**(5766): 1481–1484.

Zweig, M.H. and Campbell, G. (1993) Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**(4): 561–577.