

Library Declaration Form



University of Otago Library

Author's full name and year of birth: Simon Thomas Kelly,
(for cataloguing purposes) 24 February 1992

Title of thesis: A Bioinformatics Approach to Synthetic Lethal Interactions in Breast Cancer with Gene Expression Data

Degree: Doctor of Philosophy

Department: Department of Biochemistry

Permanent Address: 710 Cumberland Street, Dunedin, NZ

I agree that this thesis may be consulted for research and study purposes and that reasonable quotation may be made from it, provided that proper acknowledgement of its use is made.

I consent to this thesis being copied in part or in whole for

- i) a library
- ii) an individual

at the discretion of the University of Otago.

Signature:

Date:

A Bioinformatics Approach to
Synthetic Lethal Interactions in
Breast Cancer with Gene
Expression Data

S. Thomas Kelly

a thesis submitted for the degree of
Doctor of Philosophy
at the University of Otago, Dunedin,
New Zealand.

31 March 2017

Abstract

Background

Synthetic lethal interactions are re-emerging in genetics research in the genomics era driven by potential applications in precision medicine against cancers. This approach aims to exploit functional redundancy at the genetic level against mutations in cancers for developing specific treatments against them, including loss of function events in tumour suppressors. Of particular interest is the targeting loss of function of E-cadherin, encoded by *CDH1*, a tumour suppressor gene involved in Breast and Stomach cancers. Experimental screens have been used to identify candidate synthetic lethal interactions and here bioinformatics analysis used to augment the triage drug target triage process. Furthermore the pathway composition of synthetic lethal candidates and the effect of pathway structure on their detection in genomics data.

Approach

A computational statistics methodology, the Synthetic Lethal Prediction Tool (SLIPT) has been developed to detect synthetic lethal interactions in gene expression datasets. The methodology has been demonstrated on Breast and Stomach cancer datasets from The Cancer Genome Atlas (TCGA) database, testing for interactions with *CDH1*. Various analyses have been applied to further elucidate these candidates, including differential gene expression, correlation co-expression, unsupervised clustering, gene set over-representation analysis, singular-value decomposition “metagenes”, and permutation resampling analysis. A particular challenge of performing these analyses was to compare SLIPT gene candidates to the results of an experimental synthetic lethal siRNA screen of E-cadherin Telford *et al.*

(2015) at the pathway level. Graph theory methods including information centrality and shortest paths were applied to the most supported pathways from both the computational and experimental synthetic lethal candidates to test for graph structure among hits from each approach. Simulation and modelling was performed to test the statistical performance of the SLIPT methodology and further applied to datasets with simulated correlation structures, including those derived from known graph stuctures.

Findings

A vast number of genes having expression consistent with being synthetic lethal partners of *CDH1* were detected in both TCGA Breast and Stomach cancer genes. For breast cancers, these genes clustered into several distinct groups, with distinct enriched biological functions and elevated expression in different clinical subclasses such as normal-like, basal, or estrogen receptor negative samples. While the number of genes detected by both computational and experimental approaches were not significant, there was significant pathway composition in the overlapping genes. In particular $G_{\alpha i}$ signalling, cytoplasmic microfibres, and extracelluar fibrin clotting were supported by both approaches even after permutation testing. These findings are consistent with the known roles of E-cadherin in cytoskeletal or cell signalling roles and the proposed downstream targets of GPCR singalling of Telford *et al.* (2015). Many of these and related pathways were replicated in the separate stomach cancer dataset. Furthermore other candidate pathways uniquely supported by the computational predictions included regulation of immune signalling and translational elongation, both unlikely to have been detected with high dose siRNA in an isogenic cell line and these are still candidates for further testing in mouse xenograft models.

A number of approaches were adapted or developed to test whether there was a connection between synthetic lethal candidates in the graph structures of the pathways most supported by prior analyses. Network centrality measures were used to compare the importance or connectivity of genes in the pathway subnetworks but no significant difference was found between synthetic candidates and other genes within the same pathway. Another hypothesis was that computational synthetic lethal candidates would be

downstream of experimental candidates within a pathway but no evidence of directionality between the candidates was detected.

A model of synthetic lethality was developed and was sucessfully implemented to simulate gene expression datasets with known underlying synthetic lethal partners of a query gene. For small numbers of known synthetic lethal partners, the SLIPT methdology performed well respect to reciever operator characteristic curves. As the number of true partners to detect increases, the power to detect them diminishes. Increasing sample sizes, however, was able to mitigate this effect somewhat as expected. This finding was replicated in simulations up to a feasible number of human genes (20,000) with more true negatives and correlations structures. The SLIPT methdology performs similarly across these conditions and performs better than Pearson's correlation (for co-expression) or the χ^2 -test without a directional criterion. However, correlation structure of the dataset does impact on synthetic lethal predictions, genes correlated with (or in a pathway structure near to) true synthetic lethal partners having elevated test statistic values over other true negatives. A quadratic (second order polynomial) least squares linear regression methodology has been developed as a comparable alternative with the added benefit of conditioning against known partners (or strongest candidates prior analyses).

Thus my thesis has developed, evaluated and refined a bioinformatics approach to discovery of synthetic lethal genes solely from gene expression data.

Research Contributions During Candidature

Publications

Kelly, S. T. and Spencer, H. G. (2017) Population-Genetics Models of Sex-Limited Genomic Imprinting. *Theoretical Population Biology* **115**:35-44
doi:10.1016/j.tpb.2017.03.004

Manuscripts Submitted

Kelly, S. T., Single, A. B., Telford, B. J., Beetham, H. G, Godwin, T. D., Chen, A., Black, M., A., and Guilford, P. J. (2017) Towards HDGC chemoprevention: vulnerabilities in E-cadherin-negative cells identified by genomic interrogation of isogenic cell lines and whole tumors. Submitted to *Cancer Prevention Research*.

Kelly, S. T., Chen, A., Guilford, P. J., and Black, M. A. (2017) Synthetic lethal interaction prediction of target pathways in E-cadherin deficient breast cancers. Submitted to *BMC Genomics*.

Software Packages

Several software packages in the R language have been released on GitHub while preparing this thesis. Please see the appropriate GitHub repository for more information (including documentation and vignettes) on installing and running these packages, on the following account: <https://github.com/TomKellyGenetics>

- `slipt` to accompany the synthetic lethal publication above and release SLIPT (Synthetic Lethal Interaction Prediction Tool)
- `vioplotx` to provide enhanced violin plots
- `heatmap.2x` to provide annotated heatmaps
- `igraph.extensions` metapackage to provide the packages for iGraph objects:
 - `plot.igraph` to provide plotting for directed graphs
 - `info.centrality` to compute network analysis metrics
 - `pathway.structure.permutation` to perform resampling genes within on pathway structures

- `graphsim` to simulate gene expression with `mvtnorm` from pathway structures

Conference Participation

eResearch 2017 (Queenstown) Speaker February 20th-22nd “Detecting Synthetic Lethality from Cancer Gene Expression: A PhD project on genetic interactions with CDH1 inactivation in TCGA data”

RIKEN Division of Genomic Technologies (Yokohama, Japan) Seminar 2016 October 20th; National Institute of Genetics (Mishima, Japan) Seminar 2016 October 21st; Tokyo University Institute of Medical Science (Shirokanedai Campus, Japan) Seminar 2016 October 24th; Sokendai Graduate University (Hayama, Japan) Seminar 2016 October 25th “Analysis of Synthetic Lethal Pathways in Breast Cancer: A PhD project on genetic interactions with CDH1 inactivation in TCGA data”

Next Generation Sequencing Asia 2016 (Singapore) Poster October 11th-12th “Bioinformatic Investigations of Synthetic Lethal Interactions with E-cadherin in Breast Cancer” (Supported by the University of Otago Division of Health Sciences; Maurice and Phyllis Paykel Trust)

eResearch 2016 (Queenstown, New Zealand) Speaker February 9th-11th “Sifting the Needles in the Haystack: Permutation Resampling Biological Pathways in Cancer Genomic Interaction Data” (Supported by REANNZ)

Genetics Otago Symposium 2016 (Dunedin, New Zealand) Student Speaker March 7th-8th “A Bioinformatics approach to Genetic Interactions: Synthetic Lethal Pathways with E-cadherin in Breast Cancer Genomics Data”

Research Bazaar 2015 (Melbourne, Australia) February 16th-18th “My digital research toolkit” (Supported by the New Zealand eScience Infrastructure)

QMB Cancer Drugs Satellite 2014 (Queenstown, New Zealand) Poster August 24th-25th; Otago School of Medical Sciences 2015 (Dunedin, New Zealand) Poster Postgraduate Symposium April 28th-29th Bioinformatics Prioritisation of Synthetic Lethal Targets for Drug Activity Against E-Cadherin Deficient Cancers

DunDead: Zombie Science and Popular Culture Festival 2014 (Dunedin, New Zealand) Ignite Speaker August 16th-17th “Hidden in Plain Sight - The Genetics of Zombies”

eResearch 2014 (Waikato University, Hamilton, New Zealand) Ignite Speaker June 30th-July 2nd “Bioinformatic analysis of synthetic lethal genetic interactions in breast cancer” (Supported by Google)

Acknowledgements

I thank my supervisors A/Prof. Mik Black and Prof. Parry Guilford for their support and guidance throughout this my postgraduate studies. It has been a great experience, I look forward to seeing what your research groups produce in the future, may this not be the end.

I am also thankful for the guidance and mentorship of Prof. Hamish Spencer for career advice throughout my studies and time in his research group.

I am also grateful to the past and current members of these research groups, and my peers at the laboratory benches and computers across campus. The peer support, comraderie, and guidance to newer students has been an incredible part of my time at Otago and has made my thesis studies not just easier but possible at all. The postgraduate community is very special here and have truly made some lifelong friends from all over the world, you are talented researchers and amazing people. May we meet again some day. Wherever you may end up, there's always time to catch up and I'd be delighted to host some visits while working abroad.

I cannot thank my friends, flatmates and family enough for their patience and support during such as massive, challenging, and (I'm sure you've heard too many times) stressful undertaking during both my PhD and the study leading up to it. There are too many of you to name everyone here without leaving someone out, so thank you all for everything you've done, both the good times and the tough. Thank you for pretending to understand when I try to discuss complex math at the wrong moment. Thank you for checking my writing or slides, even if I should have given you more time. Thank for your time when all I really needed was a chat over a walk or a pint and a moment to think clearly.

I must also thank various organisations supported this research project:

- This thesis was supported by the Postgraduate Tassell Scholarship in Cancer Research, a University of Otago Doctoral Scholarship.
- The New Zealand eScience Infrastructure (NeSI) provided access to the Intel Pan high-performance computing cluster, support, and training to use it effectively. Various aspects of this thesis would not have been possible without access to such a resource.
- The Health Research Council (HRC) of New Zealand provided funding for experimental research in the Cancer Genetics Laboratory. Again some aspects of this project would not have been possible without access to the data and findings funded by this grant.
- The Allan Wilson Centre and Otago School of Biomedical Sciences provided funding for summer research placements which was a valuable opportunity to gain experience and training used in this thesis project.

I thank the following organisations for support towards presenting findings in this thesis at conference and seminars:

- Google (towards eResearch 2014 conference, Hamilton)
- NeSI (towards Software Carpentry training and Research Bazaar 2015, Melbourne)
- REANNZ, NZGL, and NeSI (towards eResearch 2016 conference, Queenstown)
- Otago Division of Health Sciences, Department of Biochemistry, Oxford Global, and Maurice and Phyllis Paykel Trust (towards NGS Asia 2016, Singapore)
- RIKEN Division of Genomics Technologies and the Okinawa Institute of Science and Technology (for hosting seminars in Japan)

Contents

Glossary	xxi
Acronyms	xxiii
1 Introduction	1
1.1 Cancer Research in the Post-Genomic Era	1
1.1.1 Cancer as a Global Health Concern	1
1.1.1.1 The Genetics and Molecular Biology of Cancers	3
1.1.2 The Human Genome Revolution	6
1.1.2.1 The First Human Genome Sequence	6
1.1.2.2 Impact of Genomics	7
1.1.3 Technologies to Enable Genetics Research	7
1.1.3.1 DNA Sequencing and Genotyping Technologies	7
1.1.3.2 Microarrays and Quantitative Technologies	8
1.1.3.3 Massively Parallel “Next Generation” Sequencing	9
1.1.3.3.1 Molecular Profiling with Genomics Technology .	10
1.1.3.3.2 Sequencing Technologies	11
1.1.3.4 Bioinformatics as Interdisciplinary Genomic Analysis .	12
1.1.4 Follow-up Large-Scale Genomics Projects	12
1.1.5 Cancer Genomes	13
1.1.5.1 The Cancer Genome Atlas Project	14
1.1.5.1.1 Findings from Cancer Genomes	15
1.1.5.1.2 Genomic Comparisons Across Cancer Tissues .	16
1.1.5.1.3 Cancer Genomic Data Resources	17
1.1.6 Genomic Cancer Medicine	18
1.1.6.1 Cancer Genes and Driver Mutations	18
1.1.6.2 Personalised or Precision Cancer Medicine	19
1.1.6.2.1 Molecular Diagnostics and Pan-Cancer Medicine	20
1.1.6.3 Targeted Therapeutics and Pharmacogenomics	20
1.1.6.3.1 Targeting Oncogenic Driver Mutations	21
1.1.6.4 Systems and Network Biology	21
1.1.6.4.1 Network Medicine, and Polypharmacology	24
1.2 A Synthetic Lethal Approach to Cancer Medicine	25
1.2.1 Synthetic Lethal Genetic Interactions	25
1.2.2 Synthetic Lethal Concepts in Genetics	26
1.2.3 Studies of Synthetic Lethality	27

1.2.3.1	Synthetic Lethal Pathways and Networks	27
1.2.3.1.1	Evolution of Synthetic Lethality	28
1.2.4	Synthetic Lethal Concepts in Cancer	29
1.2.5	Clinical Impact of Synthetic Lethality in Cancer	30
1.2.6	High-throughput Screening for Synthetic Lethality	32
1.2.6.1	Synthetic Lethal Screens	34
1.2.7	Computational Prediction of Synthetic Lethality	37
1.2.7.1	Bioinformatics Approaches to Genetic Interactions . .	37
1.2.7.2	Comparative Genomics	38
1.2.7.3	Analysis and Modelling of Protein Data	41
1.2.7.4	Differential Gene Expression	42
1.2.7.5	Data Mining and Machine Learning	43
1.2.7.6	Bimodality	46
1.2.7.7	Rationale for Further Development	47
1.3	E-cadherin as a Synthetic Lethal Target	47
1.3.1	The <i>CDH1</i> gene and it's Biological Functions	48
1.3.1.1	Cytoskeleton	48
1.3.1.2	Extracellular and Tumour Micro-Environment	48
1.3.1.3	Cell-Cell Adhesion and Signalling	49
1.3.2	<i>CDH1</i> as a Tumour (and Invasion) Suppressor	49
1.3.2.1	Breast Cancers and Invasion	49
1.3.3	Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer .	50
1.3.4	Models of <i>CDH1</i> loss in cell lines	51
1.4	Summary and Research Direction of Thesis	51
1.4.1	Thesis Aims	53
2	Methods and Resources	55
2.1	Bioinformatics Resources for Genomics Research	55
2.1.1	Public Data and Software Packages	55
2.1.1.1	Cancer Genome Atlas Data	56
2.1.1.2	Reactome and Annotation Data	57
2.2	Data Handling	57
2.2.1	Normalisation	57
2.2.2	Sample Triage	58
2.2.3	Metagenes and the Singular Value Decomposition	60
2.2.3.1	Candidate Triage and Integration with Screen Data .	60
2.3	Techniques	61
2.3.1	Statistical Procedures and Tests	61
2.3.2	Gene Set Over-representation Analysis	62
2.3.3	Clustering	62
2.3.4	Heatmap	63
2.3.5	Modeling and Simulations	63
2.3.5.1	Receiver Operating Characteristic (Performance) . .	64
2.3.6	Resampling Analysis	64
2.4	Pathway Structure Methods	65
2.4.1	Network and Graph Analysis	65

2.4.2	Sourcing Graph Structure Data	66
2.4.3	Constructing Pathway Subgraphs	67
2.4.4	Network Analysis Metrics	67
2.5	Implementation	68
2.5.1	Computational Resources and Linux Utilities	68
2.5.2	R Language and Packages	69
2.5.3	High Performance and Parallel Computing	72
3	Methods Developed During Thesis	74
3.1	A Synthetic Lethal Detection Methodology	74
3.2	Synthetic Lethal Simulation and Modelling	77
3.2.1	A Model of Synthetic Lethality in Expression Data	77
3.2.2	Simulation Procedure	81
3.3	Detecting Simulated Synthetic Lethal Partners	84
3.3.1	Binomial Simulation of Synthetic lethality	84
3.3.2	Multivariate Normal Simulation of Synthetic lethality	86
3.3.2.1	Multivariate Normal Simulation with Correlated Genes	89
3.3.2.2	Specificity with Query-Correlated Pathways	96
3.3.2.3	Importance of Directional Testing	96
3.4	Graph Structure Methods	98
3.4.1	Upstream and Downstream Gene Detection	98
3.4.1.1	Permutation Analysis for Statistical Significance	99
3.4.1.2	Hierarchy Based on Biological Context	100
3.4.2	Simulating Gene Expression from Graph Structures	101
3.5	Customised Functions and Packages Developed	105
3.5.1	Synthetic Lethal Interaction Prediction Tool	105
3.5.2	Data Visualisation	106
3.5.3	Extensions to the iGraph Package	109
3.5.3.1	Sampling Simulated Data from Graph Structures	109
3.5.3.2	Plotting Directed Graph Structures	109
3.5.3.3	Computing Information Centrality	110
3.5.3.4	Testing Pathway Structure with Permutation Testing	110
3.5.3.5	Metapackage to Install iGraph Functions	111
4	Synthetic Lethal Analysis of Gene Expression Data	112
4.1	Synthetic Lethal Genes in Breast Cancer	113
4.1.1	Synthetic Lethal Pathways in Breast Cancer	115
4.1.2	Expression Profiles of Synthetic Lethal Partners	116
4.1.2.1	Subgroup Pathway Analysis	119
4.2	Comparing Synthetic Lethal Gene Candidates	122
4.2.1	Primary siRNA Screen Candidates	122
4.2.2	Comparison with Correlation	123
4.2.3	Comparison with Primary Screen Viability	125
4.2.4	Comparison with Secondary siRNA Screen Validation	126
4.2.5	Comparison to Primary Screen at Pathway Level	128
4.2.5.1	Resampling Genes for Pathway Enrichment	130

4.2.6	Integrating Synthetic Lethal Pathways and Screens	133
4.3	Metagene Analysis	135
4.3.1	Pathway Expression	136
4.3.2	Somatic Mutation	138
4.3.3	Synthetic Lethal Pathway Metagenes	142
4.3.4	Synthetic Lethality in Breast Cancer	143
4.4	Replication in Stomach Cancer	144
4.5	Discussion	145
4.5.1	Strengths of the SLIPT Methodology	145
4.5.2	Synthetic Lethal Pathways for E-cadherin	146
4.5.3	Replication and Validation	148
4.5.3.1	Integration with siRNA Screening	148
4.5.3.2	Replication across Tissues	149
4.6	Summary	149
5	Synthetic Lethal Pathway Structure	151
5.1	Synthetic Lethal Genes in Reactome Pathways	151
5.1.1	The PI3K/AKT Pathway	152
5.1.2	The Extracellular Matrix	154
5.1.3	G Protein Coupled Receptors	157
5.1.4	Gene Regulation and Translation	157
5.2	Network Analysis of Synthetic Lethal Genes	158
5.2.1	Gene Connectivity and Vertex Degree	159
5.2.2	Gene Importance and Centrality	160
5.2.2.1	Information Centrality	160
5.2.2.2	PageRank Centrality	162
5.3	Relationships between Synthetic Lethal Genes	164
5.3.1	Hierarchical Pathway Structure	164
5.3.1.1	Contextual Hierarchy of PI3K	164
5.3.1.2	Testing Contextual Hierarchy of Synthetic Lethal Genes	164
5.3.2	Upstream or Downstream Synthetic Lethality	168
5.3.2.1	Measuring Structure of Candidates within PI3K	168
5.3.2.2	Resampling for Synthetic Lethal Pathway Structure .	170
5.4	Discussion	172
5.5	Summary	174
6	Simulation and Modeling of Synthetic Lethal Pathways	176
6.1	Comparing Synthetic Lethal Detection Methods	177
6.1.1	Performance of SLIPT and χ^2 across Quantiles	178
6.1.1.1	Correlated Query Genes affects Specificity	181
6.1.2	Alternative Synthetic Lethal Detection Strategies	183
6.1.2.1	Correlation for Synthetic Lethal Detection	184
6.1.2.2	Testing for Bimodality with BiSEp	185
6.2	Simulations with Graph Structures	186
6.2.1	Performance over a Graph Structure	187
6.2.1.1	Simple Graph Structures	187

6.2.1.2	Constructed Graph Structures	188
6.2.2	Performance with Inhibitions	192
6.2.3	Synthetic Lethality across Graph Structures	198
6.2.4	Performance within a Simulated Human Genome	201
6.3	Simulations in More Complex Graph Structures	206
6.3.1	Simulations over Pathway-based Graphs	207
6.3.2	Pathway Structures in a Simulated Human Genome	210
6.4	Discussion	213
6.4.1	Simulation Procedure	213
6.4.2	Comparing Methods with Simulated Data	214
6.4.3	Design and Performance of SLIPT	215
6.4.4	Simulations from Graph Structures	217
6.5	Summary	218
7	Discussion	220
7.1	Synthetic Lethality and <i>CDH1</i> Biology	220
7.1.1	Established Functions of <i>CDH1</i>	221
7.1.2	The Molecular Role of <i>CDH1</i> in Cancer	221
7.2	Significance	222
7.2.1	Synthetic Lethality in the Genomic Era	222
7.2.2	Clinical Interventions based on Synthetic Lethality	224
7.3	Evaluating the Synthetic Lethality Prediction Tool	225
7.3.1	Strength of the Synthetic Lethality Prediction Tool	225
7.3.2	Limitations of the Synthetic Lethality Prediction Tool	225
7.3.3	Comparisons to Alternative Methods	225
7.3.3.1	Combined with Experimental Screening	225
7.3.3.2	Differences to Computational Methods	225
7.4	Future Directions	225
7.4.1	Refinements Synthetic Lethality Prediction Methods	227
7.4.1.1	Wider Use of Synthetic Lethality Prediction	227
7.4.2	Validation of Synthetic Lethal Genes and Pathways	227
7.4.2.1	Pre-clinical and Clinical Testing	227
7.4.3	Application to Further Genes and Pathways	227
8	Conclusion	228
References		232
A	Sample Quality	258
A.1	Sample Correlation	258
A.2	Replicate Samples in TCGA Breast	261
B	Software Used for Thesis	265

C Mutation Analysis in Breast Cancer	274
C.1 Synthetic Lethal Genes and Pathways	274
C.2 Synthetic Lethal Expression Profiles	277
C.3 Comparison to Primary Screen	280
C.3.1 Resampling Analysis	282
C.4 Compare SLIPT genes	284
C.5 Metagene Analysis	286
C.6 Expression of Somatic Mutations	287
C.7 Metagene Expression Profiles	290
D Intrinsic Subtyping	293
E Stomach Expression Analysis	295
E.1 Synthetic Lethal Genes and Pathways	295
E.2 Comparison to Primary Screen	299
E.2.1 Resampling Analysis	301
E.3 Metagene Analysis	303
F Synthetic Lethal Genes in Pathways	307
G Pathway Connectivity for Mutation SLIPT	315
H Information Centrality for Gene Essentiality	319
I Pathway Structure for Mutation SLIPT	322
J Performance of SLIPT and χ^2	325
J.1 Correlated Query Genes affects Specificity	331
K Graph Structures	337
K.1 Simulations from Simple Graph Structures	337
K.1.1 Simulations from Inhibiting Graph Structures	339
K.2 Simulation across Graph Structures	342
K.3 Simulations from Complex Graph Structures	346
K.3.1 Simulations from Complex Inhibiting Graphs	349
K.4 Simulations from Pathway Graph Structures	356

List of Tables

1.1	Methods for Predicting Genetic Interactions	37
1.2	Methods for Predicting Synthetic Lethality in Cancer	38
1.3	Methods used by Wu <i>et al.</i> (2014)	40
2.1	Excluded Samples by Batch and Clinical Characteristics.	58
2.2	Computers used during Thesis	69
2.3	Linux Utilities and Applications used during Thesis	69
2.4	R Installations used during Thesis	70
2.5	R Packages used during Thesis	70
2.6	R Packages Developed during Thesis	72
4.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT	114
4.2	Pathways for <i>CDH1</i> partners from SLIPT	116
4.3	Pathway composition for clusters of <i>CDH1</i> partners from SLIPT	120
4.4	Analysis of variance (ANOVA) for Synthetic Lethality and Correlation with <i>CDH1</i>	124
4.5	Comparing SLIPT genes against secondary siRNA screen in breast cancer	127
4.6	Pathway composition for <i>CDH1</i> partners from SLIPT and siRNA screening	129
4.7	Pathways for <i>CDH1</i> partners from SLIPT	132
4.8	Pathways for <i>CDH1</i> partners from SLIPT and siRNA primary screen	134
4.9	Candidate synthetic lethal metagenes against <i>CDH1</i> from SLIPT	143
5.1	ANOVA for Synthetic Lethality and Vertex Degree	160
5.2	ANOVA for Synthetic Lethality and Information Centrality	162
5.3	ANOVA for Synthetic Lethality and PageRank Centrality	164
5.4	ANOVA for Synthetic Lethality and PI3K Hierarchy	167
5.5	Resampling for pathway structure of synthetic lethal detection methods	171
B.1	R Packages used during Thesis	265
C.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from mtSLIPT	275
C.2	Pathways for <i>CDH1</i> partners from mtSLIPT	276
C.3	Pathway composition for clusters of <i>CDH1</i> partners from mtSLIPT	279
C.4	Pathway composition for <i>CDH1</i> partners from mtSLIPT and siRNA	281
C.5	Pathways for <i>CDH1</i> partners from mtSLIPT	282
C.6	Pathways for <i>CDH1</i> partners from mtSLIPT and siRNA primary screen	283
C.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from mtSLIPT	286

D.1	Comparison of Intrinsic Subtypes	293
E.1	Synthetic lethal gene partners of <i>CDH1</i> from SLIPT in stomach cancer	295
E.2	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	296
E.3	Pathway composition for clusters of <i>CDH1</i> partners in stomach SLIPT	298
E.4	Pathway composition for <i>CDH1</i> partners from SLIPT and siRNA screening	300
E.5	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	301
E.6	Pathways for <i>CDH1</i> partners from SLIPT in stomach and siRNA screen	302
E.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from SLIPT in stomach cancer	303
G.1	ANOVA for Synthetic Lethality and Vertex Degree	318
G.2	ANOVA for Synthetic Lethality and Information Centrality	318
G.3	ANOVA for Synthetic Lethality and PageRank Centrality	318
H.1	Information centrality for genes and molecules in the Reactome network	320
I.1	ANOVA for Synthetic Lethality and PI3K Hierarchy	322
I.2	Resampling for pathway structure of synthetic lethal detection methods	324

List of Figures

1.1	Synthetic genetic interactions	26
1.2	Synthetic lethality in cancer	30
2.1	Read count density	59
2.2	Read count sample mean	59
3.1	Framework for synthetic lethal prediction	75
3.2	Synthetic lethal prediction adapted for mutation	76
3.3	A model of synthetic lethal gene expression	78
3.4	Modeling synthetic lethal gene expression	79
3.5	Synthetic lethality with multiple genes	80
3.6	Simulating gene function	82
3.7	Simulating synthetic lethal gene function	82
3.8	Simulating synthetic lethal gene expression	83
3.9	Performance of binomial simulations	85
3.10	Comparison of statistical performance	85
3.11	Performance of multivariate normal simulations	87
3.12	Simulating expression with correlated gene blocks	90
3.13	Simulating expression with correlated gene blocks	91
3.14	Synthetic lethal prediction across simulations	92
3.15	Performance with correlations	93
3.16	Comparison of statistical performance with correlation structure	94
3.17	Performance with query correlations	95
3.18	Statistical evaluation of directional criteria	96
3.19	Performance of directional criteria	97
3.20	Simulated graph structures	101
3.21	Simulating expression from a graph structure	103
3.22	Simulating expression from graph structure with inhibitions	104
3.23	Demonstration of violin plots with custom features	107
3.24	Demonstration of annotated heatmap	107
3.25	Simulating graph structures	110
4.1	Synthetic lethal expression profiles of analysed samples	118
4.2	Comparison of SLIPT to siRNA	122
4.3	Compare SLIPT and siRNA genes with correlation	123
4.4	Compare SLIPT and siRNA genes with correlation	124
4.5	Compare SLIPT and siRNA genes with viability	125

4.6	Compare SLIPT genes with siRNA viability	126
4.7	Resampled intersection of SLIPT and siRNA candidates	130
4.8	Pathway metagene expression profiles	137
4.9	Expression profiles for constituent genes of PI3K	139
4.10	Expression profiles for estrogen receptor related genes	140
4.11	Somatic mutation against the PI3K metagene	141
5.1	Synthetic Lethality in the PI3K Cascade	153
5.2	Synthetic Lethality in the Elastic Fibre Formation Pathway	155
5.3	Synthetic Lethality in the Fibrin Clot Formation	156
5.4	Synthetic Lethality and Vertex Degree	159
5.5	Synthetic Lethality and Centrality	162
5.6	Synthetic Lethality and PageRank	163
5.7	Hierarchical Structure of PI3K	165
5.8	Hierarchy Score in PI3K against Synthetic Lethality in PI3K	166
5.9	Structure of Synthetic Lethality in PI3K	168
5.10	Structure of Synthetic Lethality Resampling in PI3K	169
6.1	Performance of χ^2 and SLIPT across quantiles	179
6.2	Performance of χ^2 and SLIPT across quantiles with more genes	180
6.3	Performance of χ^2 and SLIPT across quantiles with query correlation .	181
6.4	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	183
6.5	Performance of negative correlation and SLIPT	185
6.6	Simple graph structures	188
6.7	Performance of simulations on a simple graph	189
6.8	Performance of simulations is similar in simple graphs	190
6.9	Performance of simulations on a pathway	191
6.10	Performance of simulations on a simple graph with inhibition	193
6.11	Performance is higher on a simple inhibiting graph	195
6.12	Performance of simulations on a constructed graph with inhibition . . .	196
6.13	Performance is affected by inhibition in graphs	197
6.14	Detection of Synthetic Lethality within a Graph Structure with Inhibitions	199
6.15	Performance of simulations including a simple graph	203
6.16	Performance on a simple graph improves with more genes	204
6.17	Performance on an inhibiting graph improves with more genes	205
6.18	Performance of simulations on the PI3K cascade	209
6.19	Performance of simulations including the PI3K cascade	211
6.20	Performance on pathways improves with more genes	212
A.1	Correlation profiles of removed samples	259
A.2	Correlation analysis and sample removal	260
A.3	Replicate excluded samples	261
A.4	Replicate samples with all remaining	262
A.5	Replicate samples with some excluded	263
C.1	Synthetic lethal expression profiles of analysed samples	278

C.2	Comparison of mtSLIPT to siRNA	280
C.3	Compare mtSLIPT and siRNA genes with correlation	284
C.4	Compare mtSLIPT and siRNA genes with correlation	284
C.5	Compare mtSLIPT and siRNA genes with siRNA viability	285
C.6	Somatic mutation against PIK3CA metagene	287
C.7	Somatic mutation against PI3K protein	288
C.8	Somatic mutation against AKT protein	289
C.9	Pathway metagene expression profiles	290
C.10	Expression profiles for p53 related genes	291
C.11	Expression profiles for BRCA related genes	292
E.1	Synthetic lethal expression profiles of stomach samples	297
E.2	Comparison of SLIPT in stomach to siRNA	299
F.1	Synthetic Lethality in the PI3K/AKT Pathway	307
F.2	Synthetic Lethality in the PI3K/AKT Pathway in Cancer	308
F.3	Synthetic Lethality in the Extracellular Matrix	309
F.4	Synthetic Lethality in the GPCRs	310
F.5	Synthetic Lethality in the GPCR Downstream	311
F.6	Synthetic Lethality in the Translation Elongation	312
F.7	Synthetic Lethality in the Nonsense-mediated Decay	313
F.8	Synthetic Lethality in the 3' UTR	314
G.1	Synthetic Lethality and Vertex Degree	315
G.2	Synthetic Lethality and Centrality	316
G.3	Synthetic Lethality and PageRank	317
H.1	Information centrality distribution	321
I.1	Synthetic Lethality and Heirarchy Score in PI3K	322
I.2	Heirarchy Score in PI3K against Synthetic Lethality in PI3K	323
I.3	Structure of Synthetic Lethality in PI3K	323
I.4	Structure of Synthetic Lethality Resampling	324
J.1	Performance of χ^2 and SLIPT across quantiles	325
J.2	Performance of χ^2 and SLIPT across quantiles	327
J.3	Performance of χ^2 and SLIPT across quantiles with more genes	329
J.4	Performance of χ^2 and SLIPT across quantiles with query correlation .	331
J.5	Performance of χ^2 and SLIPT across quantiles with query correlation .	333
J.6	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	335
K.1	Performance of simulations on a simple graph	338
K.2	Performance of simulations on an inhibiting graph	339
K.3	Performance of simulations on a constructed graph with inhibition	340
K.4	Performance of simulations on a constructed graph with inhibition	341
K.5	Detection of Synthetic Lethality within a Graph Structure	342
K.6	Detection of Synthetic Lethality within an Inhibiting Graph Structure .	344

K.7	Detection of Synthetic Lethality within an Inhibiting Graph Structure	345
K.8	Performance of simulations on a large graph	346
K.9	Performance of simulations on a branching graph	347
K.10	Performance of simulations on a complex graph	348
K.11	Performance of simulations on a large constructed graph with inhibition	350
K.12	Performance of simulations on a large constructed graph with inhibition	351
K.13	Performance of simulations on a branching graph with inhibition	352
K.14	Performance of simulations on a branching graph with inhibition	353
K.15	Performance of simulations on a complex graph with inhibition	354
K.16	Performance of simulations on a complex graph with inhibition	355
K.17	Performance of simulations on the $G_{\alpha i}$ signalling pathway	356
K.18	Performance of simulations including the $G_{\alpha i}$ signalling pathway	357

Glossary

bioinformatics	Statistical or computational approaches to biological data or research tools.
bisulfite-Seq	Methylome data from sequencing bisulfite treated DNA.
CAGE-Seq	Transcriptome data from cap analysis of gene expression.
E-cadherin	Epithelial cadherin (calcium-dependent adhesion), a cell-adhesion protein encoded by the tumour suppressor gene <i>CDH1</i> .
epigenome	An analysis of epigenetic modifications of all genes in the genome.
exome	An approach or technology designed to generate or use data from all genes in the genome.
genomics	An approach or technology designed to generate or use data from all genes in the genome.
metabolome	An analysis of all the metabolites and enzymes in the cell.
metagenomics	An analysis of all the genes and genomes in a community.
microRNA	Short RNA molecules generally regarded to regulate gene expression by binding to mRNA.
omics	A rationale for applying molecular studies across all of the genes in the genome or biomolecules in the cell.
Pan cancer	Study or analysis of cancers across tissue of origin.
pleiotropy	A gene which has multiple biological functions.

proteome	An analysis of all the proteins expressed from the genome.
RNA-Seq	Transcriptome data from sequencing RNA.
Sanger	A dideoxy chain termination method for DNA sequencing (named after Fred Sanger).
synthetic lethal	Genetic interactions where inactivation of multiple genes is inviable (or deleterious) when they are viable if inactivated separately.
transcriptome	An analysis of all the genes expressed in the genome.

Acronyms

AMP	Adenosine monophosphate.
AMPK	AMP-activated protein kinase.
ANOVA	Analysis of Variance.
AUROC	Area under the receiver operating characteristic (curve).
BioPAX	Biological Pathway Exchange.
BiSEp	Bimodal Subsetting Expression.
BMP	Bone morphogenic protein.
CCLE	Cancer Cell Line Encyclopaedia.
cDNA	Complementary deoxyribonucleic acid (from mRNA).
ChIP-Seq	Chromatin immunoprecipitation sequencing.
COSMIC	Catalogue Of Somatic Mutations In Cancer.
CXCR	Chemokine receptors.
DAISY	Data mining synthetic lethal identification pipeline.
DDBJ	DNA Data Bank of Japan.
DNA	Deoxyribonucleic acid.
EMBL	European Molecular Biology Laboratory.
EMT	Epithelial-mesenchymal transition.
ENA	The European Nucleotide Archive.
GEO	Gene Expression Omnibus.
GO	Gene Ontology.
GPCR	G protein coupled receptor.
HDAC	Histone deacetylase.
HDGC	Hereditary diffuse gastric cancer.
HLRCC	Hereditary leiomyomatosis and renal cell carcinoma.

HPC	High Performance Computing.
ICGC	International Cancer Genome Consortium.
InDel	Insertion or deletion (in DNA sequence).
JAK	Janus kinase.
lncRNA	Long non-coding ribonucleic acid.
miRNA	microRNA.
mRNA	Messenger ribonucleic acid.
NCBI	National Center for Biotechnology Information (in the USA).
NCI	National Cancer Institute (in the USA).
NeSI	New Zealand eScience Infrastructure.
NGS	Next-generation sequencing.
NHGRI	National Human Genome Research Institute (in the USA).
NIG	National Institute of Genetics (in Japan).
NIH	National Institutes of Health (in the USA).
NMD	Nonsense-mediated decay.
PAM50	Prediction Analysis of Microarray 50.
PCR	Polymerase chain reaction.
PDE	Phosphodiesterase.
PI3K	Phosphoinositide 3-kinase.
PIP ₂	Phosphatidylinositol (4,5)-bisphosphate.
PIP ₃	Phosphatidylinositol (3,4,5)-trisphosphate.
qPCR	Quantitative (real-time) polymerase chain reaction.
RFLP	Restriction fragment length polymorphism.
RGS	G-protein signalling.
RHO	Ras Homolog Family.
RNA	Ribonucleic acid.
RNAi	Ribonucleic acid interference.
ROC	Reciever operating characteristic (curve).
RRBS	Reduced representation bisulfite sequencing.

SGI	synthetic genetic interaction.
siRNA	Short interfering ribonucleic acid.
SLIPT	Synthetic lethal interaction prediction tool.
Slurm	Simple Linux Utility for Resource Management.
SNP	Single nucleotide polymorphism.
TCGA	The Cancer Genome Atlas (genomics project).
TGF β	Transforming growth factor β .
UCSC	University of California, Santa Cruz.
UTR	Untranslated region (of mRNA).
WNT	Wingless-related integration site.

Chapter 1

Introduction

1.1 Cancer Research in the Post-Genomic Era

Genomics technologies have the potential to vastly impact upon various areas including health and cancer medicine. Considering the progress in recent genomics research, this technology and the findings from it have considerable potential for significant impacts in the clinic and wider applications of genetics either directly or by enabling more focused genetics research on candidates selected from genomics or bioinformatics analysis. The completion of the draft Human Genome (*Lander et al.*, 2001) marks a major accomplishment in genetics research and raises new challenges to utilise this genomic scale information effectively. Technologies in this area have rapidly developed since completion of the human genome project and many global large-scale projects have expanded upon the human genome, to populations (1000 Genomes, 2010), to cancers (Dickson, 1999; Zhang *et al.*, 2011), and to deeper functional understanding (Kawai *et al.*, 2001; ENCODE, 2004). However, impact on the clinic has been slower than initially anticipated following the completion of the “draft” genome with genomics technologies yet to become widely adopted in healthcare and oncology. Here we outline the genomics technologies and bioinformatics approaches which have led to availability of genomics data and techniques used in this thesis and potential for applications in cancer research or the clinic in the future.

1.1.1 Cancer as a Global Health Concern

Cancer is a class of diseases involving malignant cellular growth, invasion of tissues, and spread to other organs. While there are also environmental factors, most cancers occur more frequently with age and family history. Accordingly, genetics has become widely acknowledged as having an important role in cancer risk (in addition to environmental

factors). Cancers arise from dysregulated cellular growth or differentiation from stem cells. These can occur through genetic mutations or alterations in gene regulation or expression.

Cancers are a major global health concern, being the second leading cause of death globally (WHO, 2017), with an estimated annual incidence of 14.1 million cases and annual mortality of 8.2 million people (Ferlay *et al.*, 2015). Breast and stomach cancers are among the 5 most frequent cancers globally, with breast cancer affecting women more than other cancer tissue types. Breast cancer has an estimated annual incidence of 1.6 million cases and mortality of 520 thousand people. Stomach cancer has an estimated annual incidence of 950 thousand cases and a mortality of 723 thousand people. Cancer is also a major health concern here in New Zealand, with 19.1 thousand people (including 2.5 thousand cases of breast cancer and 370 cases of stomach cancer) diagnosed annually (Hanna, 2003), among the highest incidence (age-standardised per capita) of cancer in the world (Ferlay *et al.*, 2015).

While the genetic contribution to cancer risk and many of the molecular changes occurring in cancers are widely acknowledged (ASCO, 2017; Cancer Research UK, 2017; Cancer Society of NZ, 2017), the majority of these findings have yet to impact on clinical practice. Diagnostics are traditionally based on pathological examination of tissue samples where histological staining for cell type, biomolecules and biomarkers continue to be widely used, although genetic and biochemical markers are being adopted for some cancer types. In general, the current standard of care includes surgery, radiation, and cytotoxic chemotherapy, depending on whether the cancer is localised or has become systemic (via metastasis) and spread to other organ systems. These approaches can be effective against certain cancers, particularly in early stage cancer or in patients with particular subtypes (such as acute myeloid leukaemia) which respond well to modern treatment regimens. Thus early diagnosis is important to patient survival and quality of life. National screening programs (which prioritise patients with a high risk of cancer) therefore aim to diagnose cancers early and subtypes more accurately. Therefore identification of patients with genetic variants or family histories (for inclusion in national databases) for high risk of particular cancers is an important health issue, particularly where effective interventions exist if these cancers are diagnosed early. Thus high risk individuals being regularly monitored for some cancers are sometimes offered preventative surgery or treatment for pre-cancerous tissue (Guilford *et al.*, 2010; Scheuer *et al.*, 2002).

Chemotherapy is a last-resort treatment for many advanced stage (systemic) cancers which is designed to inhibit the growth and spread of cancer throughout the body by targeting rapidly growing cells. However, this approach often has severe adverse effects, a narrow therapeutic window, and is not suitable for chemopreventative application in many cases (Kaelin, Jr, 2009). Since surveillance preventative surgery (removing the tissue at risk of cancer) is not completely effective at preventing cancers and may impact on quality of life, depending on the cancer tissue types they are at risk of, alternative treatment strategies based on molecular biology and other fields are being investigated. These alternatives include immunological, endocrine, and targeted therapeutics, with a particular interest in treatments with specificity against cancer cells and wider applications (i.e., tolerable effective doses in applications as a chemopreventative or against advanced stage cancers).

1.1.1.1 The Genetics and Molecular Biology of Cancers

Cancers involve dysregulation of genes with both somatic mutations or regulatory disruptions which accumulate during a patient's lifetime and germline mutations which predispose individuals to high-risk early onset cancers (American Cancer Society, 2017; Guilford *et al.*, 1998; NCI, 2015). Cancer is widely viewed to be a genetic disease due to these familial cancer syndromes, hereditary risk factors, and the molecular changes occurring in cancers, including numerous cancer genes which have been identified Stratton *et al.* (2009); Vogelstein *et al.* (2013). Cancer genes are generally classified into two classes: "oncogenes" which are activated in cancers, driving tumour growth and invasion, or "tumour suppressors" which are inactivated in cancers, removing cellular regulation and genomic maintenance functions. The mutations which cause cancers accumulate with age and have been suggested to be inevitably coupled with aging due to the association of cancer incidence with the stem cell divisions in which mutations could occur across tissue types (Tomasetti and Vogelstein, 2015).

Hanahan and Weinberg (2000) identified several key molecular and cellular traits shared across most cancers as a rational approach to the complex changes that occur in cancer initiation and progression due to common molecular machinery underlying all cells. A cancer cell must possess limitless replication potential, modulate growth signals to grow indefinitely, and gain invasive or metastatic capabilities. In addition, cancers must evade apoptosis, the immune system, and sustain angiogenesis and energy metabolism in order to survive (Hanahan and Weinberg, 2000, 2011). In order to achieve this, cancer cells undergo changes to their genomes and the surrounding cells to create a tumour microenvironment. Thus genomic instability has a key role in the

survival and proliferation of cancer cells and the progression of further disease, as these malignant characteristics are acquired. Identifying the mechanisms of these acquired traits and the underlying genetic mutation or dysregulation behind them, such as E-cadherin mutation in metastasis or p53 mutation in genomic instability (Hanahan and Weinberg, 2000), will be an important step in understanding and inhibiting cancer with the next generation of genomically-informed treatments.

Molecular biological processes have particular importance in characterising breast cancers. Gene expression and regulatory signals confer cell identity and response to the environment. Therefore gene expression has been investigated with microarray technologies Perou *et al.* (2000), with “intrinsic subtypes” identified characterised by estrogen receptor, *HER2*, and basal, epithelial signalling. The expression profiles were similar across independent samples of the same tumour and between primary and metastatic tumours of the same patient. Thus expression profiles represent the molecular state of a tumour rather than the sample and the molecular configuration of the cells regulation is carried through the cellular lineage of during metastasis preserving the molecular subtype. These molecular intrinsic subtypes “luminal A”, “luminal B”, “*HER2*-enriched”, “basal-like”, and “normal-like” have been replicated across microarray studies (Hu *et al.*, 2006), with their relevance to prognosis (including predicting survival and response to neoadjuvant chemotherapy) demonstrated and a 50-gene subtype predictor from microarray and quantitative PCR (qPCR) analysis has been provided (Parker *et al.*, 2009; Sørlie *et al.*, 2001). This has been further updated with the “claudin-low” subtype (Herschkowitz *et al.*, 2007) and stimulated further investigations into subtyping of breast cancers by molecular properties.

Despite differences in subtyping performed by different research groups and companies, there is widespread agreement that distinguishing luminal, *HER2*-enriched, and triple negative tumours can be performed with expression profiles and have value in our understanding of cancer progression and prognostic importance for patients Dai *et al.* (2015). High-throughput technologies have the potential to enable such subtyping on a vast scale in discovery of further subtypes in breast cancer or other diseases and in identification of these subtypes along with mutations in routine clinical diagnostic and prognostic testing. The “Pan cancer” approaches by the cancer genome atlas project (as discussed in more detail in Section 1.1.5.1.1) expand on the importance of molecular differences between cancers by examining molecular profiles across cancer tissue types (Weinstein *et al.*, 2013).

The molecular variability of cancer has also been approached rationally at a pathway level with patients subgroups activating different molecular pathways reflecting differences in disease mechanisms (Gatza *et al.*, 2010). A robust approach to measuring pathway activation in cancer is with a “metagene” which gives a consistent signal as a consensus of expression across genes even if they are inversely correlated (Anjomshoaa *et al.*, 2008; Huang *et al.*, 2003; Nagalla *et al.*, 2013). These are derived from the first principal component or eigenvector of a singular matrix decomposition, capturing the the most consistent variation across genes in a pathway or gene signature. Gatza *et al.* (2010) used gene signatures for 18 cellular pathways in breast cancer to define subtypes with distinct molecular pathway activity. In a meta-analysis of Affymetrix microarray expression data for 1,143 samples and 50 cell lines, unsupervised hierarchical clustering robustly defined subtypes with common homogeneous pathway activity despite variation in the specific mutations giving rise to them. These subtypes with shared pathway activity have similar molecular characteristics (such as DNA copy number), clinical properties and prognosis, building upon the intrinsic subtypes (Parker *et al.*, 2009) and providing a functional interpretation for molecular stratification (Gatza *et al.*, 2010). The pathway-based subtypes often correspond to intrinsic subtypes (Gatza *et al.*, 2010, 2014) and provide finer molecular stratification such as environmental stress response (to hypoxia within HER2-enriched cancers) (Gatza *et al.*, 2011) and have pathway-specific DNA copy number variation or essential genes (Gatza *et al.*, 2014). Gatza *et al.* (2014) extend these analyses include 52 pathway signatures from previous publications in breast cancer, replicating known characteristics (such as hormone re) of each subtype and identifying novel subtype-specific driver genes of proliferation by analysis of microarray expression from 476 from The Cancer Genome Atlas (TCGA). In addition to distinct biological functions driving growth of breast cancer subtypes, these molecular subtypes provide a rational approach based on molecular properties to cancer treatment with combination and targeted therapies (Gatza *et al.*, 2010, 2014; Hanahan and Weinberg, 2000).

Cancer is a major health concern with a well-established genetic contribution, in risk and in the molecular changes occurring during progression (Stratton *et al.*, 2009). Many genes have been discovered to be important in different cancers with molecular differences between cancers, including alterations across the genome, being of clinical importance. As such cancers were among the first samples investigated with genomics following the sequencing of the human genome Dickson (1999) and continue to be the subject of genomics and bioinformatics investigations.

1.1.2 The Human Genome Revolution

The advent of the Human Genome sequence (Lander *et al.*, 2001) has transformed genetics research including the study of health and disease (Lander, 2011; Peltonen and McKusick, 2001). Systematic, unbiased studies across all of the genes in the genome are viable in unprecedented ways. The successful undertaking of such an international scientific megaproject has set an example for numerous initiatives to follow, including many genomics investigations expanding to species, to the functional, or to the population level (Collins *et al.*, 2003). These projects serve as an excellent resource for genetics research globally, particularly for cancers where genomics investigation have been widely applied to different tissues across molecular profiles Bamford *et al.* (2004); Weinstein *et al.* (2013); Zhang *et al.* (2011) . Genome sequencing technologies continue to improve, decrease in price, and become increasingly feasible in a wider range of research and clinical applications.

1.1.2.1 The First Human Genome Sequence

The first human genome is a good example of a large-scale genomics project for it's success as an international collaboration and releasing their data as a resource for the wider scientific community (Collins *et al.*, 2003; Lander *et al.*, 2001). This particular project generated significant public interest due to it being a landmark achievement, the first of it's scale, and some controversial findings. Namely, the number of genes discovered (particularly those specific to vertebrates) was much lower than most estimates for a genome of it's size and the number of repetitious transposon elements was very high. Even the figure of 30–40,000 genes given by the original publication is now regarded to be an overestimate (Ezkurdia *et al.*, 2014; IHGSC, 2004).

Accounting for the “complexity” encoded by the human genome with so few genes has led to investigations into molecular function, expression profiling, and population variation. When announcing the draft genome, Lander *et al.* (2001) conceded that genomic information alone was not sufficient for biological understanding and that many investigations remained to be carried out, with their objective being to share the raw genome data so that it was available for further inquiry rather than interpreting it themselves. While genomics technologies and genomics projects have flourished since then, the need in turn for systematic means of interpreting data of such scale and for the interdisciplinary expertise to do so has only grown.

1.1.2.2 Impact of Genomics

Genomics has stimulated investigations into many of these previously largely explored areas of functional genetics and thus been of immense value in genetics research, attracting high expectations for further applications. Genomics research created widespread anticipation for potential applications in healthcare, agriculture, ecology, conservation, and evolutionary biology despite few of these being realised yet.

Cancer research is an area of particularly high expectations for the clinical impact of genomics in oncology. Genomics technologies have potential applications across cancer diagnostics, prognosis, management, and developing treatment. Cancers often involve genetic mutation or dysregulated gene expression which can be detected in a genome or transcriptome with potential to improve patient care. While direct impact of genomics on the clinic has been limited, compared to initial expectations following the publication of the human genome, diagnostic cancer genes and therapeutic targets identified with genomics research have begun to be introduced in the clinic (Stratton *et al.*, 2009).

1.1.3 Technologies to Enable Genetics Research

1.1.3.1 DNA Sequencing and Genotyping Technologies

Genotyping was once commonly performed on variable regions of the genome with restriction fragment length polymorphism (RFLP) or repetitious microsatellite regions. These exploited sequence variation at target sites of restriction enzymes or measured the length of repetitious regions, using polymerase chain reaction (PCR), restriction enzymes, and gel electrophoresis to measure deoxyribonucleic acid (DNA) genotypes at particular sites. This is laborious and limited to well characterised variable regions of the genome, generally genes or nearby marker regions.

The Sanger (dideoxy) chain termination method (Sanger and Coulson, 1975) enabled DNA sequencing and genotyping at a widespread scale, being less technically difficult than the Maxam-Gilbert sequencing by degradation method (Gilbert and Maxam, 1973; Maxam and Gilbert, 1977), which required more radioactive and toxic reactants. The Sanger methodology has relatively long read length (particularly compared to early versions of more recent technologies), with read lengths of 500–700 base pairs accurately sequenced in most applications, usually following targeted amplification with PCR. Sanger sequencing by gel electrophoresis takes around 6–8 hours and has been further refined with the “capillary” approach to 1–3 hours and requiring less input DNA and reactants. The capillary approach has been scaled up to run in parallel from

a 96 well plate, at 166 kilobases per hour. The 96 well parallel capillary method was one of the main innovations which made the first Human Genome Project feasible and was used throughout (Lander *et al.*, 2001). Due to the quality of the Sanger sequence reads and low cost, it is still widely used in smaller scale applications, clinical testing, and to validate the findings of newer approaches.

1.1.3.2 Microarrays and Quantitative Technologies

Real-time or qPCR is another adaptation of genetic technologies to quantitatively study nucleic acids, often reverse transcribed “deoxyribonucleic acid (cDNA)” or messenger “messenger ribonucleic acid (mRNA)” to measure (relative) gene expression or transcript abundance. While numerous quality control measures are required to correctly interpret a qPCR experiment, these have similarly become widely adopted and are still used for smaller scale experiments and as a “gold standard” for measuring gene expression (Adamski *et al.*, 2014). This also represents a shift in the application of qPCR and sequencing technology, where the primary interest is quantifying the amount of input material (by the rate of amplification to a certain level) rather than the qualitative nature of the sequence itself. The more recent technologies of microarrays and RNA-Seq have similarly embraced this application to quantify DNA copy number, ribonucleic acid (RNA) expression, and DNA methylation levels. Due to results of comparable or arguably better quality from these newer technologies (Beck *et al.*, 2016; Git *et al.*, 2010; McCourt *et al.*, 2013; Robin *et al.*, 2016), this “gold standard” status has begun to come under scrutiny.

Microarrays represent a truly high-throughput molecular technique, reducing the cost, time, and labour required to study molecular factors such as genotype, expression, or methylation across many genes, making it feasible to do so over a statistically meaningful number of samples. Microarrays are manufactured with probes which measure binding of particular nucleotide sequences to either quantitatively detect the presence of a sequence such as a single nucleotide polymorphism (SNP) or quantify DNA copy number, gene expression, or DNA CpG methylation. Microarray technologies have popularised “genome scale” studies of genetic variation and expression.

In addition to being more versatile and higher-throughput than PCR based techniques, microarrays are considered cost-effective, particularly when scaled up to a large number of probes. They are also available with established gene panels or customised probes from a number of commercial manufacturers. These remained popular during the introduction of newer technologies due to reliability and relatively lower cost, especially in large-scale projects involving many samples. However, microarrays have

issues with signal-to-noise ratio, with both sensitivity to low nucleic acid abundance and “saturation” of probes at high abundance, edge effects, and requiring more starting material than qPCR. Thus qPCR is still used for many small gene panel studies.

1.1.3.3 Massively Parallel “Next Generation” Sequencing

Similar to microarrays, the introduction of massively parallel sequencing technologies have further expanded the availability of high-throughput molecular studies to researchers, with corresponding availability of genomics data from these studies. This “Next-Generation Sequencing” (NGS) expands not only gene expression studies (compared to microarrays) but extends to genome sequencing *de novo* for previously unknown genome and transcriptome sequences at an unprecedented scale. This has been a particularly important technological revolution in genomics, as the cost and time of genome sequencing has dropped dramatically and enabled sequencing projects of far more samples and applications beyond the Human Genome Project. Particularly, when dealing with variants in a species with an existing reference sequence such as humans, where there is a low computational cost of mapping to a reference compared to a genome assembly. However, the cost of sequencing (RNA-Seq) for gene expression or DNA methylation studies is still considerably higher than a microarray study (limiting feasible sample sizes).

Compared with arrays, NGS studies have additional challenges, particularly regarding large data and compute requirements to handle the raw output data. Compared with the established methods to analyse microarray data, handling NGS data can be more technically difficult. While methods developed for analysing microarray data can be repurposed for sequence analysis in many cases, more bioinformatics expertise is required particularly to handle the raw read data and changing approaches for various developments in sequencing technologies. One of the main computational challenges is the assembly of reads or mapping to a reference genome due to the inherently small reads of most NGS technologies compared to the Sanger methodology. Furthermore, there are fewer software releases and best practices established specifically for RNA-Seq data, thus many analyses are still conducted with customised analysis approaches and command-line tools. Compared to existing graphical tools or pipelines for microarray analysis, this is a more active technology for bioinformatics research where many applications of genomics data have yet to be explored.

However, the methodology itself has challenges with the sample preparation, requiring a relatively high quantity of input material and “contamination” with over abundant ribosomal rRNA taking up the majority of the sequencing if not purified

correctly. This abundance of rRNA is a particularly important issue in microarray and RNA experiments in Eukaryotes where it is commonplace to target the mRNA by binding to the poly-A tail (RNA-Seq) or 5' cap (CAGE-Seq). However, this has the potential to exclude miRNA (miRNA) and long non-coding ribonucleic acid (lncRNA) of interest unless the sample is prepared specifically to study these. Similarly capturing a subsection of the genome for exome analysis or reduced representation bisulfite sequencing (RRBS), focuses on sequencing DNA sequences and methylation levels of CpG sites near known genes to reduce cost, noise, and incidental findings.

In many cases, the benefits of NGS technologies over microarrays still outweigh the additional cost. NGS technologies have the advantage of greater potential accuracy and sensitivity than microarrays, depending on the sequencing depth or “coverage”, theoretically sensitive down to the exact number of molecules for each transcript. NGS experiments are regarded as “reproducible” with no need for technical replicates, although these are still performed for a subset of samples in many projects for quality assurance purposes. NGS has a wider dynamic range than microarrays and is able to detect SNPs, insertions or deletions, and splice variants in addition to quantifying DNA copy number or transcript abundance. NGS scales to all genes and beyond for these molecular applications without having to design new probes as required for a microarray. Thus NGS technologies are not limited to genes with already characterised sequence or functions, do not need to be updated with new probes for each genome annotation release, and do not require a reference genome at all for new species. A “transcriptome” can be assembled *de novo* for an expression study in any organism by sequencing the mRNA extracted from a cell.

1.1.3.3.1 Molecular Profiling with Genomics Technology

NGS is highly adaptable to different applications: DNA sequencing (whole genome or exome), DNA methylation (bisulfite-Seq), RNA-Seq, miRNA, lncRNA, or chromatin immunoprecipitation sequencing (ChIP-Seq). RNA-Seq of the transcriptome is a common adaptation where RNA is reverse transcribed and sequenced from the resulting cDNA. This is utilised to quantify the levels of RNA and identify which regions of DNA are expressed. Similar bisulfite treatment converts cytosine residues to uracil (sequenced as thymidine), sparing methylated cytosine enabling it to be distinguished with bisulfite-Seq for high-throughput detection of the notable epigenetic mark and is a common procedure to generate an epigenome. Subsets of the nucleic acid may be extracted for sequencing such as the coding regions of DNA (for the “exome”), the

mRNA 5' cap (CAGE-Seq), mRNA 3' poly-A tail (RNA-Seq), microRNA, or an enriched subset of variable regions for DNA sequencing (“genotyping by sequencing”) and methylation studies (“reduced-representation bisulfite sequencing). High-throughput gel and mass spectrometry techniques have been applied to proteins and metabolites to generate the proteome and metabolome respectively. These “omics” technologies are applicable across a wide range of biomolecules in a cell and these “molecular profiles” are produced in many experimental laboratories.

1.1.3.3.2 Sequencing Technologies

Illumina sequencing (developed by Solexa and later acquired by Illumina) was released in 2006. It utilises reversible terminating dyes to sequence by synthesis with a lower accuracy (98%) and read lengths of 150–250bp. Illumina more than makes up for relatively short reads (along with improving the read length of the technology) and low accuracy with high-throughput and cost effectiveness, with a Hi-Seq 4000 platform producing up to 10 billion paired-end reads (1500Gbp) in a run of appropriately 3 days, capable of sequencing 12 human genomes ($30\times$ coverage) or 100 human transcriptomes simultaneously (Illumina, 2017). Illumina has further reduced the cost of sequencing with the economies of scale with the Hi-Seq X 10 claiming to produce a human genome (with $30\times$ coverage) for less than US\$1000, the first platform to achieve this long-standing goal in genomics. The high-throughput of Illumina sequencing also makes deep sequencing for high coverage, high quality consensus reads, and sensitive RNA-Seq experiments feasible. Illumina sequencing now has a dominating market share of the NGS technologies. Their Hi-Seq platforms were used in large-scale genomics projects (such as the cancer genome atlas discussed in Section 1.1.5.1) to generate the sequence data used throughout this thesis.

NGS technologies continue to be refined with Illumina (and competitors) continuing to release refined productions, offer additional genomics-based services, and decreases overhead and operating costs to enable a wider range of research projects. As such RNA-Seq for examining the transcriptome of an organisms and for expression studies in diseases is a growing field of research and expression data will continue to be generated for a range of samples in many healthy and diseased tissues. The technology be also be further improved developments in speed and accuracy (such as Ion Torrent platforms) and towards long reads of single molecules (such as Pacific Biosciences, Oxford Nanopore, and Quantum Biosystems Japan).

Due to such benefits of sequencing over previous technologies (and their continued refinement), this thesis has focused on gene expression data generated by RNA-Seq rather than microarrays. RNA-Seq data is widely available as a resource from large-scale cancer genomics projects and methods to make inferences from RNA-Seq experiments could feasibly be applied to many other studies based on these current (or similar future) technologies.

1.1.3.4 Bioinformatics as Interdisciplinary Genomic Analysis

Genomics technologies have given rise to data at a scale previously rarely encountered in molecular biology, making inference with conventional techniques difficult. Computational, Mathematical, and Statistical skills are required to handle this data effectively, in addition to biological background to frame and interpret research questions. Drawing upon these disciplines to handle biological data has become the field of “Bioinformatics”, focusing specifically on making inferences from genomics and high-throughput molecular data or developing the tools to do so. This contrasts with the existing fields of “theoretical” or “computational biology” which existed prior to genomics data, focusing on modelling and simulating aspects of biology without necessarily addressing the genomics data or detecting the phenomena in nature, extending beyond genetics to cell modelling, neuroscience, cancer development, ecology, and evolution.

In practice, many researchers identify with both bioinformatics and computational biology, or draw upon the findings and methods of the other field. This thesis uses many approaches in bioinformatics to biological research questions and established mathematical or bioinformatics resources.

Gene expression analysis is the focus of many bioinformatics research groups, drawing upon statistical approaches to appropriately handle microarray and RNA-Seq data along with making biological inferences from a large number of statistical tests. This presents various challenges from normalising sample data and accounting for batch effects to developing or applying statistical tests tailored to biological hypotheses and testing them at a genome-wide scale, generally across thousands of genes. There are numerous approaches for dealing with these challenges, some of which will be described in Chapter 2.

1.1.4 Follow-up Large-Scale Genomics Projects

A number of projects have attempted to follow up on the human genome project to varying degrees of success. The genomes have since been sequenced for a variety of model organisms, organisms of importance in health, agriculture, metagenomics of

microorganisms (microbiome), ecology and conservation. Genomics projects have also been applied functional genetics (Kawai *et al.*, 2001; ENCODE, 2004) and to human populations with an interest variability between individuals and health or disease risk (HapMap, 2003; 1000 Genomes, 2010).

Genomics databases have also focused on facilitating distribution of genomic data generated by researchers, rather than generating it themselves. GenBank hosted by the National Center for Biotechnology Information (NCBI) in the US, the European Nucleotide Archive (ENA) hosted by the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ) hosted by the National Institute of Genetics (NIG) do so by serving as public repositories of DNA sequence data. Gene Expression Omnibus (GEO) (Clough and Barrett, 2016), arrayExpress (Rustici *et al.*, 2013), and caArray (Heiskanen *et al.*, 2014) serve a similar purpose as a resource for gene expression datasets, originally developed for microarray data but RNA-Seq data is now supported by some platforms. They are repositories for researchers to deposit, share, and access gene expression data, serving as a resource to support ongoing research where larger datasets than would were previously accessible for many individual laboratories are available (Rung and Brazma, 2013). These resources cover not only DNA sequence across the genome but also molecular profiles of other factors by adapting genomic sequencing or other high throughput technologies for quantifying gene expression or DNA methylation. Sharing the expression datasets generated in a publication is now required by some journals.

Similarly, international projects and consortiums have begun to release data gathered using common agreed upon protocols in laboratories across the world, often hosting public databases of these themselves, publishing their own investigations into the datasets as they are released, or offering basic searches and analytics of the data via a web portal. These databases include many of the genomics projects discussed above and the cancer-specific projects discussed below. In many ways, the quality, consistency, and accessibility of these international projects has become more appealing than accessing smaller studies, particularly for gene expression datasets where the more recent, larger projects have switched from microarray to RNA-Seq technologies. This distinction will also be discussed later.

1.1.5 Cancer Genomes

The importance of genomics technologies in the future of cancer research was noticed, even in the early days of genomics (Dickson, 1999). The Cancer Genome Project

(CGP) based at Wellcome Trust Sanger Institute in the UK were among the first to launch investigations into cancer after the publication of the Human Genome, using this genome sequence, consensus across the cancer research literature, and sequencing the genes of cancers themselves. Initially, the Sanger Institute set out to sequence 20 genes across 378 samples while the Human Genome project was still ongoing (Collins and Barker, 2007), optimising sequencing and computation infrastructure for a larger project while doing so. The main aim of the Cancer Genome Project was to discover “cancer genes”, those frequently mutated in cancers by comparing the genes of cancer and normal tissue samples, both “oncogenes” and “tumour suppressors” which are activated and inactivated respectively in cancers. This project is ongoing and the UK continues to be involved in international sequencing initiatives and those focused on particular tissue types.

The Sanger Institute also hosts the Catalogue of Somatic Mutations in Cancer (COSMIC, 2016), a database and website of cancer genes. This launched with 66,634 samples and 10,647 mutations from initial investigations into *BRAF*, *HRAS*, *KRAS*, and *NRAS* (Bamford *et al.*, 2004). It has since expanded to include 1,257,487 samples with 4,175,8787 gene mutations curated from 23,870 publications, including 29,112 whole genomes (COSMIC, 2016). This database now also identifies cancer genes from DNA copy number, differential gene expression and differential DNA methylation.

1.1.5.1 The Cancer Genome Atlas Project

Based in the US, The Cancer Genome Atlas (TCGA) project was established in 2005, a combined effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) (TCGA, 2017a). They first set out to demonstrate the pilot project on brain (McLendon *et al.*, 2008), ovarian (Bell *et al.*, 2011), and squamous cell lung (Hamerman *et al.*, 2012) cancers. In 2009, the project expanded aiming to analyse 500 samples each for 20-25 tumour tissue types. They have since exceeded that goal, with data available for 33 cancer types including 10 “rare” cancers, a total of over 10,000 samples.

The TCGA projects set out to generate a molecular “profile” of the tumour (and some matched normal tissue) samples: the genotype, somatic mutations, gene expression, DNA copy number, and RNA methylation levels. While these were originally performed largely with microarray technologies, exome and RNA-Seq has been since adopted and performed for many TCGA samples, with whole genomes being performed for some samples. Data which cannot be used to identify the patients (such as somatic mutation, expression, methylation, and various clinical factors) are publicly available.

1.1.5.1.1 Findings from Cancer Genomes

The Cancer Genome Atlas pilot projects (Bell *et al.*, 2011; Hammerman *et al.*, 2012; McLendon *et al.*, 2008) serve to demonstrate the power of applying genomics technologies to cancer research at such as scale. In addition to sequencing the whole genome or a subset (exome), DNA copy number, gene expression, DNA methylation, and somatic mutations were also analysed. The initial projects used microarray technologies for expression and methylation data but these have since been replaced by RNA-Seq for expression. TCGA demonstrated the potential discovery of the molecular basis of cancer by analysing 206 glioblastoma brain cancer samples (McLendon *et al.*, 2008), highlighting the roles of *ERBB2*, *NF1*, *TP53*, and *PIK3R1* mutations, along with altered methylation of *MGMT*, and the core pathways of RTK, p53, and RB signalling in brain cancer. An analysis of 489 serious ovarian cancers (Bell *et al.*, 2011) similarly reported *TP53* mutations specifically over-represented in high grade tumours and reported 133 copy number variants, 168 differentially methylated regions, and recurrently somatic mutations in 9 genes in low grade tumours including *NF1*, *BRCA1*, *BRCA2*, *RB1*, and *CDK12*. Four transcriptional subtypes of ovarian cancers were identified, alterations in *BRCA1*, *BRCA2*, and *CCL6* had an impact on patient survival, and the homologous recombination, NOTCH and FOXM1 signalling pathways were involved in ovarian cancer growth. The genomics of 178 squamous cell lung cancers (Hammerman *et al.*, 2012) were highly complex, averaging at 360 mutations in coding regions. While no targeted therapies existed for this cancer subtype, 11 recurrently mutated genes were identified including *TP53* and *HLA-A*. The pathways altered in various squamous cell lung cancers were NFE2L2, KEAP1, differentiation genes, PI3K, CDKN2A and RB1. These aberrant genes and pathways represent potential therapeutic targets which could be identified for most samples.

The TCGA breast cancer analysis (TCGA, 2012) consisted of 802 samples with exomes, copy number variants, RPPA protein quantification, and DNA methylation, mRNA, and microRNA arrays with 97 whole genomes sequenced. Four main molecular classes were identified to subtype the samples, despite considerable heterogeneity between samples. Recurrent mutations across more than 10% of samples were identified in *TP53*, *PIK3CA*, and *GATA*. TCGA further suggests subtypes by HER2 and EGFR protein levels. In a further analysis of 817 breast cancer samples including 127 invasive lobular breast and 88 mixed type samples (Ciriello *et al.*, 2015), 3 molecular subtypes of lobular breast cancer were identified. Lobular breast cancer was also characterised by recurrent mutations in *CDH1*, *PTEN*, *TBX2*, and *FOXA1*.

TCGA reported results of colon and rectal cancers in a combined analysis of 267 samples (Muzny *et al.*, 2012), finding no genomic distinction between colorectal cancers. Apart from 16% of hypermutated colorectal cancers, the remaining samples were very similar at the molecular level with 24 significantly recurrently mutated genes identified. These include the expected *APC*, *TP53*, *SMAD4*, *PIK3CA*, and *KRAS* genes. Additionally, novel recurrent mutations were identified in *ARID1A*, *SOX9*, and *FAM123* along with recurrent copy number alterations in *ERBB2* and *IFG2*. Thus the molecular findings of colon and rectal tumours can be applicable across colorectal cancers, including the known characteristics of microsatellite instability (MSI) and CpG island methylator phenotype (CIMP) found in some colorectal tumours.

The TCGA stomach cancer analysis of 295 samples (Bass *et al.*, 2014) identified 4 molecular subtypes of stomach cancers characterised by: the Epstein-Barr virus, MSI, genomics instability, and chromosomal instability. Abberations in *PD-L1*, *PIK3CA*, and *JAK2* were also identified in stomach cancers which may present therapeutic targets.

1.1.5.1.2 Genomic Comparisons Across Cancer Tissues

TCGA have identified various genes as recurrent, driver mutations across cancer types which are likely to have a role in driving the proliferation of these cancers and present a molecular target that could be applied across tissue types. These include *TP53* (in brain, lung/head/neck squamous cell, breast, colorectal, uterine, and endometrial cancers), *ERBB2/HER2/NEU* (in brain, breast, colorectal, bladder, and lung cancers), *PIK3CA*, *PIK3R1* (in brain, breast, colorectal, endometrial, bladder, clear cell renal, and lung cancers), *BRCA1/BRCA2* (in breast and ovarian cancers), *NF1* (in brain, ovarian, and skin cancers), *ARID1A* (in colorectal, endometrial, and clear cell renal cancers), *KRAS* (in colorectal, endometrial, and skin cancers), *BRAF* (in colorectal, thyroid, and skin cancers), *EGFR* (in brain, breast, and lung cancers), and *PTEN* (in breast, endometrial, and uterine cancers) (Agrawal *et al.*, 2014; Akbani *et al.*, 2015; Bass *et al.*, 2014; Bell *et al.*, 2011; Burk *et al.*, 2017; Cherniack *et al.*, 2017; Ciriello *et al.*, 2015; Collisson *et al.*, 2014; Creighton *et al.*, 2013; Hammerman *et al.*, 2012; Kandoth *et al.*, 2013; Lawrence *et al.*, 2015; McLendon *et al.*, 2008; Muzny *et al.*, 2012; TCGA, 2012; Weinstein *et al.*, 2014). In addition to disregarding the tissue-based distinction between colon and rectal cancers based on molecular similarity (Muzny *et al.*, 2012), the TCGA project have observed differences within tumour types and proposed molecular subtyping for breast, clear cell renal, papillary renal, stomach,

skin, bladder, and prostate cancers (Abeshouse *et al.*, 2015; Akbani *et al.*, 2015; Bass *et al.*, 2014; Ciriello *et al.*, 2015; Creighton *et al.*, 2013; Hammerman *et al.*, 2012; Linehan *et al.*, 2016; Muzny *et al.*, 2012; TCGA, 2012; Weinstein *et al.*, 2014).

The “Pan cancer” project (Hoadley *et al.*, 2014; Weinstein *et al.*, 2013) analysed 3527 samples across 12 tissue types for DNA, RNA, protein, and epigenetic molecular profiles. This project was initiated in 2012 to perform a comprehensive analysis of molecular data across cancer types to identify molecular similarities and differences. Recurrent *TP53* mutations characterised high grade tumours across breast, ovarian, and endometrial cancers. HER2 was identified in brain, endometrial, bladder, and lung cancers, in addition to the known role of HER2 in breast cancers. *BRCA1* and *BRCA2* mutations were also detected across cancers, mainly breast and ovarian cancers as expected. Microsatellite instability characterised both endometrial and colorectal cancers. The Pan cancer project (Hoadley *et al.*, 2014) has identified 11 molecular subtypes across these tissues, 5 of corresponding to tissue cancer types and the remainder reassigned due to molecular similarities shared across cancer types. Squamous cell lung, head, and neck and a subset bladder cancers were grouped together by molecular similarities, characterised by a high frequency of *TP53* mutations. Conversely, bladder cancers were divided into 3 of these molecular subtypes with distinct profiles. This project further supports the genomic stratification of patients, demonstrated in breast cancer (Parker *et al.*, 2009; Pereira *et al.*, 2016; Perou *et al.*, 2000), which may apply to other cancer types and to molecular characteristics across them targeting recurrent mechanisms of cancer growth and progression (Hanahan and Weinberg, 2000, 2011).

1.1.5.1.3 Cancer Genomic Data Resources

While the findings from the TCGA projects themselves are a considerable contribution to understanding cancer biology within and across tissue types, the main eventual benefit of such projects will be the availability of the data for the research community to analyse further and use to inform future investigations (McLendon *et al.*, 2008; TCGA, 2017a; Weinstein *et al.*, 2013). These serve as a vast resource of common and rare cancer types and are publicly available for further analysis (cBioPortal, 2017; TCGA, 2017a; Zhang *et al.*, 2011). This also applies to the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project which focuses on breast cancer which also aimed to identify novel molecular subtypes (Curtis *et al.*, 2012). They performed an analysis of 2433 breast cancer samples with long-term clinical data, gene expression, copy number variants, and 173 genes sequenced which identified 40 driver

mutations in breast cancer in addition to further support for molecular subtyping to identify patient groups with different clinical outcomes (Pereira *et al.*, 2016).

1.1.6 Genomic Cancer Medicine

There is much anticipation in cancer research for genomics technologies to have a clinical impact in cancer medicine: from diagnosis and prognosis to treatment developments and strategies. These may result either from direct use of genome or RNA-Seq in clinical laboratories or indirectly from biomarkers and treatments developed with research facilitated by genomics. This second strategy is likely to have a more immediate patient benefit due to the cost of genome sequencing, particularly considering adoption in public healthcare systems with a limited budget.

1.1.6.1 Cancer Genes and Driver Mutations

There are two main categories of “cancer genes” (Futreal *et al.*, 2001). Oncogenes are those activated in cancers either by gain of function mutations in proto-oncogenes, amplification of DNA copies, or elevated gene expression. Their normal functions are typically to regulate stem cells or to promote cellular growth and recurrent mutations are typically concentrated to particular gene regions. Conversely, tumour suppressor genes are those inactivated in cancer either by loss of function mutations, deletion of DNA copies, repression of gene expression, or hypermethylation. Their normal functions are typically to regulate cell division, DNA repair, and cell signalling.

Detecting these cancer genes is a major challenge in cancer biology and has been revolutionised by genomic technologies. Recurrent mutations, or DNA copy number variants and differential gene expression or DNA methylation are all indicative of cancer genes (Mattison *et al.*, 2009), which can be detected in genomics data (Pereira *et al.*, 2016; Weinstein *et al.*, 2013). Important “driver” cancer genes (Stratton *et al.*, 2009) are difficult to detect from “passenger” mutations due to patient variation, tumour heterogeneity, and genomic instability. However, many cancer genes have been replicated from previous studies or well supported from genomics data. There remains the challenge of translating the identification of cancer genes to patient benefit with characterisation of variants of unknown significance, which mutation or gene expression markers can be used to monitor tumour progression or treatment response, and design of therapeutic intervention against many molecular targets for which they have yet to be developed or repurposed from other diseases to cancers.

Driver mutations can be identified by whether they co-occur or are mutually exclusive with mutations in other genes in cancers, are recurrently mutated across a

significant proportion of samples for a specific tissue type, or if mutations are recurrent across different cancer tissue types (cBioPortal, 2017; Pereira *et al.*, 2016; COSMIC, 2016; Weinstein *et al.*, 2013; Zhang *et al.*, 2011). Approximately 140 driver mutations have been identified, including many novel genes in particular cancers from genomics studies, with 2–8 in typically occurring in each tumour usually affecting cell fate, survival, or genome maintenance (Vogelstein *et al.*, 2013).

1.1.6.2 Personalised or Precision Cancer Medicine

The notion of using a patient’s genome to tailor healthcare to an individual has been appealing since the advent of genomics, popularised with the term “personalised medicine”. This approach was expected to span from preventative lifestyle advice to effective treatments. Personalised medicine was described in contrast with current strategies of health advice, screening, prognostics, and treatments based on what works well for the majority of the population. Adverse effects of these treatments occur in a significant subpopulation, particularly demographics under-represented in clinical studies.

The importance of genomics is still recognised in translational cancer research. Applications are particularly emphasised in molecular diagnosis, prognosis, and treatments of patients already presenting with cancers in the clinic rather than preventative medicine. This is in part due to the vast number of variants of unknown clinical significance, the ethical issue of reporting on incidental findings, and the regulatory issues direct-to-consumer genetics companies have encountered offering health risk assessment.

More recently the term “genomic medicine” has been preferred to describe the paradigm of treating cancers by their genomic features, particularly grouping patients by the mutation, expression, or DNA methylation profiles of their cancers. Radical proponents advocate for these molecular subtypes to supersede tissue or cell type specific diagnosis of cancers. However, in practice they are often used in combination, with clinical and pathological factors being informative of prognosis and the medical expertise required for treatment. The related term of “precision medicine” also stems from this trend with the rationale to target these molecular subtypes with separate treatment strategies, particularly in developing and applying treatments targeted against a particular mutation specific to cancers. To this end much research in this field is focused on identifying mutations and gene expression signatures amenable to distinguishing cancers, particularly oncogenic driver mutations, and developing treatments against them.

1.1.6.2.1 Molecular Diagnostics and Pan-Cancer Medicine

There is growing support for the use of molecular tools such as mutations or gene expression signatures to diagnose tumour subtypes addition to (or in lieu of) tissue of origin or histology. This is particularly important in breast cancer where analysis of molecular data detected several distinct “intrinsic subtypes” with differences in malignancy and patient outcome which were distinguished by molecular mechanisms rather than tissue or cellular phenotype (Parker *et al.*, 2009; Perou *et al.*, 2000). Conversely, common molecular mechanisms may be shared between cancers across tissue types as discovered by the “Pan cancer” studies, such as those conducted by the TCGA and International Cancer Genome Consortium (ICGC) projects, which combined molecular profiles across tissue types Weinstein *et al.* (2013). The molecular subtypes could feasibly be included in clinical testing as a panel of biomarkers for diagnostics and prognosis. Such biomarkers also have the potential to monitor drug response or risk of recurrence. This is also raises the need for development of treatments that target these molecular subtypes.

1.1.6.3 Targeted Therapeutics and Pharmacogenomics

Targeted therapies with specificity against a molecular target are emerging as precision cancer medicine. Molecular targets can be tested in laboratory conditions with RNA interference or pharmacological agents. Identification of molecular targets is important for developing novel anti-cancer treatments along with validation and drug testing. For oncogenic mutations, the recurrent mutant variant or overexpressed gene is directly inhibited using structure-aided drug design or compound screening. However, oncogenes with high homology to other genes or tumour suppressor genes (where lost in cancers) are not amenable to direct targeting (Kaelin, Jr, 2009).

Despite controversy over their prohibitively high cost (PHARMAC, 2016), targeted therapeutics have been applied as monoclonal antibodies against oncogenes (such as *HER2*) with relative success in clinical trials (Miles, 2001), generating considerable interest in wider application of this approach. Targeted therapeutics have potential to have applications across cancer tissue types, specificity against tumour cells, wide therapeutic windows, and combination therapies (even in advanced disease or as a chemopreventative in high-risk individuals).

1.1.6.3.1 Targeting Oncogenic Driver Mutations

Oncogene targeted therapies have also been developed with some examples of effective clinical application against cancers. However, they have already begun to manifest problems with resistance, recurrence, tissue specificity, and design of inhibitors specific to oncogenic variants rather than proto-oncogene precursors. Targeted anticancer therapeutics can exploit complex interactions to distinguish normal and cancerous cells which may benefit from studies of gene regulation or interaction networks. The unexpected synergy between inhibitors of the oncogenes $BRAF^{V600E}$ and $EGFR$ in colorectal cancer is an example of such a system Prahallad *et al.* (2012).

Despite successful application of vemurafenib against $BRAF^{V600E}$ in melanomas Dienstmann and Tabernero (2011); Ravn and Matalka (2012), colorectal cancers with $BRAF^{V600E}$ mutations have poor prognosis and lack drug response. Prahallad *et al.* (2012) used an RNA interference (RNAi) screen and found that $EGFR$ inhibition is synergistic with vemurafenib against $BRAF^{V600E}$ in colon cell lines and xenografts due to feedback activation of $EGFR$. Vemurafenib induced rapid reactivation of MAPK/ERK signalling via $EGFR$ in colorectal cell lines in a tissue-specific manner Corcoran *et al.* (2012) and may be relevant to acquired resistance in melanoma Sun *et al.* (2014). Thus combination therapies against several molecular pathways may be necessary to anticipate acquired resistance Ravn and Matalka (2012) and targeted therapeutics may be further refined from understanding the pathway structure and functional interactions across cancer cells.

1.1.6.4 Systems and Network Biology

It is also important to consider that driver mutations in oncogenes and tumour suppressor genes do not occur in isolation. The genetic interaction, regulatory and cellular signalling, and metabolic reactions are all inter-related and may each be perturbed by aberrations in gene function occurring in cancers. These relationships can be represented by biological networks by mapping pairs of genes with a particular relationship. Due to the complexity of a cell, these molecular networks are very large consisting of thousands of nodes comprised by genes or proteins.

The properties of large networks were first studied by constructing random networks by randomly linking a fixed number of nodes (Erdős and Rényi, 1959, 1960). Despite the random nature of these networks, properties such as their connectivity were well characterised. The vertex degree (number of partners for each node) of random network follows a Poisson distribution, however this property does not hold in nature, suggesting

that natural networks are non-random or not formed in this way Barabási and Oltvai (2004).

This work formed the foundation for studying complex networks (van Steen, 2010), which model features of real world networks not found in Erdős and Rényi's random networks (Erdős and Rényi, 1959, 1960). The small world property, made popular by findings in social networks (Travers and Milgram, 1969), is the remarkably short path lengths between any nodes in a small world network. A small world network is well-connected with a characteristic path length (the average length of shortest paths between all pairs of nodes) proportional to the logarithm of the number of nodes. Watts and Strogatz (1998) developed a model of random rewiring of a regular network to construct random networks with the small world property and a high clustering coefficient. While these properties are more representative of networks occurring in nature, their model is limited by the degree distribution which converges to a Poisson distribution as it is rewired Barrat and Weigt (2000).

The vertex degree distribution of naturally occurring networks often follows a power law distribution with the majority of nodes having far fewer connections than average and a small subset of highly connected network 'hubs' Barabási and Albert (1999). Hubs further differentiate into 'party' hubs (which interact simultaneously with many partners) and 'date' hubs (which interact with different partners in different conditions) Han *et al.* (2004). Network hubs can also be classed as associative or dissociative depending on whether they tend toward or away from connecting directly to other network hubs (van Steen, 2010). The associative and dissociative properties can also be used to test whether nodes of a particular subgroup (e.g., gene function) associate with each other.

Barabási and Albert (1999) constructed a network model in an entirely different way to randomly generate scale-free networks which have a power law degree distribution. They constructed random networks by preferential attachment, modelling growth of a network by sequentially adding nodes with links to existing nodes. The scale-free nature of the random networks was ensured by adding new nodes with an increasing probability of attachment to an existing node if it has higher degree. These networks successfully capture the scale-free nature of many real world networks with short characteristic path length and low eccentricity resulting in super small worlds Barabási and Albert (1999). Scale-free networks are limited by a low clustering coefficient and lack of modular structure; however, they have enabled the study of scale-free network topology and

served as a basis for modified scale-free models (Dorogovtsev and Mendes, 2003; Holme and Kim, 2002).

Han *et al.* (2004) observed dynamic modularity in biological networks and suggested the network structure may underpin genetic robustness and plasticity. They focus on network hubs which are more likely to be essential genes and define the subgroups of hubs based on correlation of gene expression with protein-protein interaction partners: ‘party’ hubs (which interact simultaneously with many partners) and ‘date’ hubs (which interact with different partners in different conditions). Party and date hubs occurred most frequently within and between network modules respectively. Party hubs were considered local regulators, whereas date hubs were considered important to network connectivity as global regulators. This distinction between classes of network hubs was supported by differences in tissue specificity and clinical relevance as a proposed predictor of clinical outcome in breast cancer with an area under the receiver operating characteristic (AUROC) of 0.784 Taylor *et al.* (2009). However, correlation between expression and protein interactions were not robustly reproduced. The importance of date hubs has been criticised for assuming a bimodal distribution and basing the global importance of data hubs on a small subset Agarwal *et al.* (2010). As an alternative interpretation, (Agarwal *et al.*, 2010) suggest the importance of interactions rather than network hubs as interactions important to the network were between functionally similar proteins. Network hubs can also be classed as associative or dissociative depending on whether they tend toward or away from connecting directly to other network hubs (van Steen 2010). The associative and dissociative properties can also be used to test whether nodes of a particular subgroup (e.g., gene function) associate with each other.

Applications of network theory are diverse, including uses in social sciences, engineering, and computer science. Due to their complexity and difficulty of gathering sufficient empirical data, biological applications of network theory are relatively unexplored. High-throughput technologies such as siRNA screens, two-hybrid screens, microarrays and massively parallel sequencing have made generating genome-scale molecular data feasible and enabled analysis of biological networks at the molecular level. Many types of inter-related molecular networks can be constructed and analysed, depending on the biological application. Genetic interaction networks will be the focus of this project because they are relatively unexplored compared to other molecular networks, have potential for applications in drug discovery (particularly cancer treatment), and may lead to better understanding of the role of genetics in cellular function and disease. Genetic interactions are usually studied at a high-throughput scale in simple model or-

ganisms such as bacteria, yeasts or the nematode worm; studies in humans, mammals, and non-model organisms (where applications would have the most societal impact) are limited by cost, time and labour constraints. Computational approaches with effective predictive models are the only feasible approach to study the connectivity of a biological network in a complex metazoan cell at the genome-scale.

1.1.6.4.1 Network Medicine, and Polypharmacology

Molecular networks are biological networks consisting of biological molecules including genes, transcripts (with non-coding and microRNAs), or proteins related by known interactions and gene regulatory or metabolic pathways. Targeted therapeutics have had some success for drug discovery, particularly in anticancer applications, including exploiting these molecular networks by designing combination therapies and applying a network pharmacology framework Hopkins (2008). Rational design of drugs selective to a single target has often failed to deliver clinical efficacy. Many existing effective drugs modulate multiple proteins, having been selected for biological effects or clinical outcome rather than molecular targets. Proponents of network biology and polypharmacology (specific binding to multiple targets) recommend developing drugs with a desired target profile designed for the target topology Barabási and Oltvai (2004); Hopkins (2008). Multi-target treatments aim to achieve a clinical outcome through modulation of molecular networks since the genetic robustness of a cell often compensates for loss of a single molecular target.

While multi-target drugs may be more difficult to design, they are faster to test clinically than drug combinations which are usually required to be tested separately first Hopkins (2008). Synthetic lethal treatments for cancer, drug combinations and multi-target drugs to combat resistance to chemotherapy and antibiotics can be informed by biological networks Barabási and Oltvai (2004); Hopkins (2008). Further optimisation of timing and dosing of drug combinations may increase efficacy of treatments with low efficacy applied separately. Low doses and drug holidays are other counter intuitive approaches which may increase clinical efficacy, reduce adverse effects, and reduce drug resistance (Sun *et al.*, 2014; Tsai *et al.*, 2012).

A molecular map of the interactions and pathways in the mammalian cellular network has the potential to impact upon drug design and clinical practice, particularly in treatment of cancer and infectious disease. Characterisation of the target system and impact of existing treatments, such as *BRAF*^{V600E} and *EGFR* inhibitors, enable wider application of the mechanisms for such interventions exploiting genetic interactions or

pathways. This could lead to development of more effective treatment interventions for these systems and prediction of similar molecular systems for development of novel drug targets and combinations.

1.2 A Synthetic Lethal Approach to Cancer Medicine

Synthetic lethality has vast potential to improve cancer medicine by expanding application of targeted therapeutics to include inactivation of tumour suppressors and genes that are difficult to target directly. Synthetic lethal interactions are also studied for gene function and drug mode-of-action in model organisms. This section introduces the concept of synthetic lethality as it was originally conceived and how it has been adopted conceptually in cancer research. Detecting these interactions at scale and interpreting them is the focus of this thesis, hence we start with an overview of the concepts involved, initial work on the interaction, and the rationale for applications to cancer. Specific investigations into synthetic lethality in cancer, detection by experimental screening, and prediction by computational analysis will then be reviewed.

1.2.1 Synthetic Lethal Genetic Interactions

Genetic interactions are a core concept of molecular biology, discovered among earliest investigations of Mendelian genetics, and receiving revived interest with new technologies and potential applications. Biological epistasis is the effect of an allele at one locus “masking” the phenotype of another locus (Bateson and Mendel, 1909). Statistical epistasis is where there is significant disparity between the observed and expected phenotype of a double mutant, compared to the respective phenotypes of single mutants and the wild-type (Fisher, 1919). Fisher’s definition lends itself to quantitative traits and more broadly encompasses synthetic genetic interactions (synthetic genetic interactions). These have become popular for studies in yeast genetics and cancer drug design (Boone *et al.*, 2007; Kaelin, Jr, 2005).

Synthetic genetic interactions are substantial deviations of growth or viability from the expected null mutant phenotype (of an organism or cell) assuming additive (deleterious) effects of the single mutants. The double mutant does not necessarily have either single mutant phenotype (as shown for cellular growth phenotypes in Figure 1.1). Most SGIs are more viable than either single mutant or less viable than the expected double mutant. Mutations are “synergistic” in negative SGIs with more deviation from the wild-type than expected. Formally, “synthetic sick” (SSL) and “synthetic lethal” (SL) interactions are negative SGIs giving growth inhibition and inviability respectively.

Synthetic lethality in cancer research more broadly describes any negative SGI with specific inhibition of a mutant cell, including SSL interactions. Mutations are “alleviating” in positive SGIs with less deviation from the wild-type than expected. For viability, “suppression” and “rescue” are positive SGIs giving at least partial restoration of wild-type growth from single mutants with growth impairment and lethal phenotypes respectively. Negative SGIs were markedly more common than positive SGIs in a number of studies in model systems Boucher and Jenna (2013); Tong *et al.* (2004).

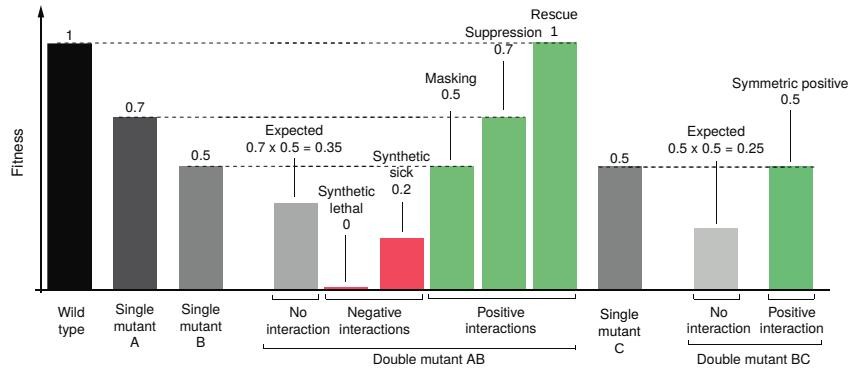


Figure 1.1: Synthetic genetic interactions. Impact of various negative and positive SGIs: negative interactions involve deleterious (sick) or inviable (lethal) phenotypes whereas positive interactions involve restoring viability by masking or suppressing the other mutation or complete rescue of the wildtype phenotype. Figure adapted from (Costanzo *et al.*, 2011) concerning growth viability fitness in yeast.

1.2.2 Synthetic Lethal Concepts in Genetics

Synthetic lethal genes are generally regarded to arise due to functional redundancy. Due to the functional level of SGIs, synthetic lethal genes do not need directly interact, nor be expressed in the same cell or at the same developmental stage: serving related functions is sufficient to affect cell (or organism) viability and be relevant to drug-mode-of-action cancer biology. Combined loss of genes performing an essential or important function in a cell are therefore deleterious. Synthetic lethal gene pairs are therefore pairwise essential with “induced essentiality”: each synthetic lethal gene becomes essential to the cell upon loss of the other.

Since synthetic lethal gene partners can be affected by extracellular stimuli such as chemicals, essentiality of synthetic lethal genes can be induced by the environment of a cell. An environmental stress condition may inhibit one or the other synthetic lethal gene, such as exposure to chemicals, in which case the synthetic lethal partner gene is

“conditionally essential” (Hillenmeyer, 2008). Thus the evolutionary rationale for the abundance of SGIs (compared to the surprisingly low number of essential genes) in a Eukaryotic genome attributed to genetic functional redundancy and network robustness of a cell which are advantageous to survival.

Biological functions are typically performed by a pathway of genes (or their products). Many genes of the same pathway may be functionally interchangeable as synthetic lethal partners of a particular gene since loss of the pathway is deleterious without the synthetic lethal partner gene. Therefore biological pathways can be subject to induced essentiality under loss of a gene and synthetic lethality may occur at pathway level or in a gene regulation network.

1.2.3 Studies of Synthetic Lethality

Genetic high-throughput screens have identified unexpected, functionally informative, and clinically relevant synthetic lethal interactions; including synthetic lethal partners of genes recurrently mutated in cancer or attributed to familial early-onset cancers. While screening presents an appealing strategy for synthetic lethal discovery, computational approaches are becoming popular as an alternative or complement to experimental methods to overcome inherent bias and limitations of experimental screens. An array of recently developed computational methods (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015; Tiong *et al.*, 2014; Wang and Simon, 2013; Wappett, 2014) show the need for synthetic lethal discovery in the fundamental genetics and translational cancer research community. However, existing computational methods are not suitable for queries of genomic data for interacting partners of a particular gene: they have been applied pairwise across the genome, do not have software released to apply the methodology, or lack statistical measures of error for further analysis. A robust prediction of gene interactions is an effective and practical approach at a scale of the entire genome for ideal translational applications, analysis of biological systems, and constructing functional gene networks.

1.2.3.1 Synthetic Lethal Pathways and Networks

SGIs are very common in genomes, with $4\times$ more interactions detected with synthetic gene array mating screens than protein-protein interactions yeast-2-hybrid studies (Tong *et al.*, 2004). The SGI network is scale-free with power-law vertex degree distribution and low average shortest path length (3.3) as expected for a complex biological network (Barabási and Oltvai, 2004). Highly connected “hub” genes with the highest number of links (vertex degree) are functionally important with many negative

SGI hubs involved in cell cycle regulation and many positive SGI hubs involved in translation (Baryshnikova *et al.*, 2010b; Costanzo *et al.*, 2010). Negative SGIs were far more common than positive SGIs, with synthetic gene loss being more likely to be deleterious to cell than advantageous which indicates than synthetic lethality may be comparably easier to detect than other SGIs.

Essential pathways are highly buffered with $5\times$ more interactions than other SGIs, consistent with strong selection for survival, as found with conditional and partial mutations in essential genes (Davierwala *et al.*, 2005). This SGI network had scale-free topology and rarely shared interactions with the protein-protein interaction network. These networks are related by an “orthogonal” relationship: shared partners in one network tend to be themselves connected directly in the other network. Essential genes were likely to have closely related functions, whereas non-essential networks were relatively more inclined to have SGIs between distinct biological pathways.

1.2.3.1.1 Evolution of Synthetic Lethality

There is poor conservation of specific SGIs between *S. cerevisiae* and *S. pombe* with 29% of the interactions tested in both distantly related species being conserved between them (Dixon *et al.*, 2008). The remaining interactions show high species-specific differences; however, many of the species-specific interactions were still conserved between biological pathways, protein complexes, or protein-protein interaction modules. Similarly, conservation of pathway redundancy was also found between Eukaryotes (*S. cerevisiae*) and prokaryotes (*E. coli*) (Butland *et al.*, 2008). Negative SGIs were more likely to be conserved between biological pathways, whereas positive SGIs were more likely to be conserved within a pathway or protein complex (Roguev *et al.*, 2008).

A modest 5% of interactions were conserved between unicellular (*S. cerevisiae*) and multicellular (*C. elegans*) organisms. However, the nematode SGI network had similar scale-free topology and modularity despite differences in methodology: metazoan RNAi screens are incomplete knockouts whereas screening null mutations is feasible in yeast (Bussey *et al.*, 2006). The nematode SGI screen identified network hubs with important interactions to orthologues of known human disease genes (Lehner *et al.*, 2006). Despite the lack of direct conservation of SGIs between yeasts and nematode worms, genetic redundancy at the gene or pathway level may yet be consistent with an induced essentiality model of SGIs where gene functions are conserved with network restructuring over evolutionary change (Tischler *et al.*, 2008). While nematode models are more closely related to human cells, cancer cells can present growth and viability

phenotypes more comparable to yeast models. Therefore findings from both SGA and RNAi models are relevant to understanding cellular network structure and in healthy and cancerous human cells. RNAi has also been applied to human and mouse cancer cells in cell culture and genetic screening experiments. These findings suggest that SGI network “rewiring” is a concern for identifying specific synthetic lethal interactions in cancer and a pathway approach may be more robust in the context of evolution, patient variation, tumour heterogeneity, and disease progression.

1.2.4 Synthetic Lethal Concepts in Cancer

Loss of function occurs in many genes in cancers including tumour suppressors and yet few interventions target such mutations compared to targeted therapies for gain of function mutation in oncogenes (Kaelin, Jr, 2005). Synthetic lethality is a powerful design strategy for therapies selective against loss of gene function with potential for application against a range of genes and diseases (Fece de la Cruz *et al.*, 2015; Kaelin, Jr, 2009). Since synthetic lethality affects cellular viability by indirect functional relationships between genes, it is suitable for indirectly targeting mutations in cancers. Once synthetic lethal partners of cancer genes are identified, targeted therapeutics can be applied against them. When genes are disrupted in cancers, the induced essentiality of synthetic lethal partners is a vulnerability that may be exploited for anti-cancer therapy. This has the potential to be very specific against cancer cells (with the target mutation) over non-cancer cells (with a functional compensating gene). Analogous to “oncogene addiction”, where cancer cells adapt to particular oncogenic growth signals and become reliant on them to remain viable (Luo *et al.*, 2009; Weinstein, 2000), synthetic lethal partners of inactivated tumour suppressors are required to maintain cancer cell viability and proliferation. As such cancers are subject to “non-oncogene addiction” and are feasible anti-cancer drug targets.

The synthetic lethal approach to cancer medicine is most amenable to loss of function mutations in tumour suppressor genes, where it would feasibly be effective against any loss of function mutation across the tumour suppressor with a viable synthetic lethal partner gene (as shown in Figure 1.2). However, the approach may also be suitable for cases where cancer cells have mutations where the normal function of the gene is disrupted such as if it were overexpression (“synthetic dosage lethality”) or if an oncogene interfered with the function of the proto-oncogenic variant such as competitive inhibition. Thus synthetic lethality expands the range of cancer-specific mutations feasible to target with targeted therapeutics to absence of tumour suppressor genes and

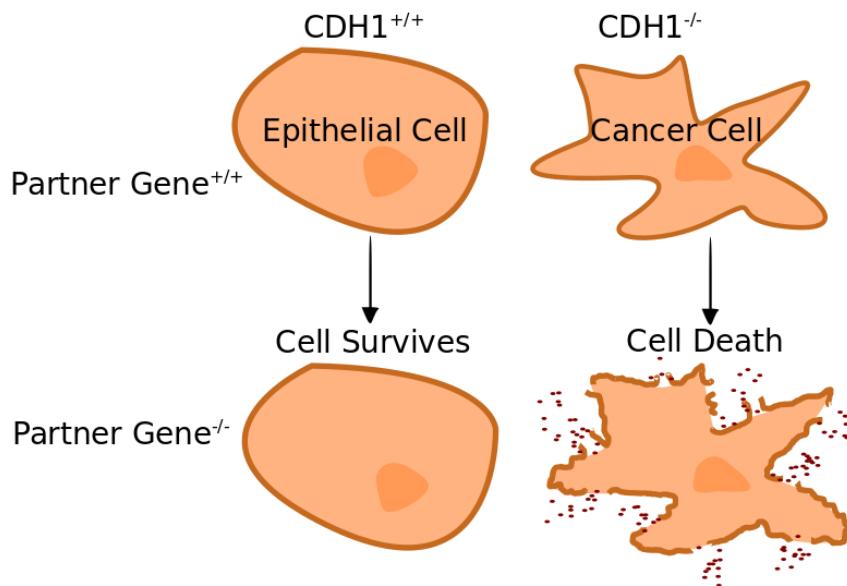


Figure 1.2: **Synthetic lethality in cancer.** Rationale of exploiting synthetic lethality for specificity against a tumour suppressor gene (e.g., *CDH1*) while other cells are spared under the inhibition of a partner gene.

distinguishing highly homologous oncogenes by functional differences by targeting their synthetic lethal partners.

1.2.5 Clinical Impact of Synthetic Lethality in Cancer

The synthetic lethal interaction of *BRCA1* or *BRCA2* with *PARP1* in breast cancer is an example of how gene interactions are important in cancer, including translation to the clinic. These genetic interactions enable specific targeting of mutations in *BRCA1* or *BRCA2* tumour suppressor genes with PARP inhibitors by inducing synthetic lethality in breast cancer (Farmer *et al.*, 2005). PARP inhibitors are one of the first targeted therapeutics against a tumour suppressor mutation with success in clinical trials.

BRCA1 or *BRCA2* and *PARP1* genes demonstrate the application of the synthetic lethal approach to cancer therapy Ashworth (2008); Kaelin, Jr (2005). *BRCA1* and *BRCA2* are homologous DNA repair genes, widely known as tumour suppressors; mutation carriers have substantially increased risk of breast (risk by age 70 of 57% for *BRCA1* and 59% for *BRCA2*) and ovarian cancers (risk by age 70 of 40% for *BRCA1* and 18% for *BRCA2*) (Chen and Parmigiani, 2007). The *BRCA1* or *BRCA2* genes, which usually repair DNA or destroy the cell if it cannot be repaired, have inactivating somatic mutations in some familial and sporadic cancers. Poly-ADP-ribose polymerase

(PARP) genes are tumour suppressor genes involved in base excision DNA repair. Loss of PARP activity results in single-stranded DNA breaks. However, *PARP1*^{-/-} knockout mice are viable and healthy indicating low toxicity from PARP inhibition (Bryant *et al.*, 2005).

Bryant *et al.* (2005) showed that *BRCA2* cells were sensitive to PARP inhibition by siRNA of *PARP1* or drug inhibition (which targets *PARP1* and *PARP2*) using Chinese hamster ovary cells, MCF7 and MDA-MB-231 breast cell lines. This effect was sufficient to kill mouse tumour xenografts and showed high specificity to *BRCA2* deficient cells in culture and xenografts. Farmer *et al.* (2005) replicated these results in embryonic stem cells and showed that *BRCA1* cells were also sensitive to PARP inhibition relative to the wild-type with siRNA and drug experiments in cell culture and drug activity against *BRCA1* or *BRCA2* deficient embryonic stem cell mouse xenografts. They found evidence that PARP inhibition causes DNA lesions, usually repaired in wild-type cells, which lead to chromosomal instability, cell cycle arrest, and induction of apoptosis in *BRCA1* or *BRCA2* deficient cells. Therefore, the pathways cooperate to repair DNA giving a plausible mechanism for combined loss as an effective anti-cancer treatment.

Thus PARP inhibitors have potential for clinical use against *BRCA1* or *BRCA2* mutations in hereditary and sporadic cancers (Ashworth 2008; Kaelin 2005). PARP inhibition has been found to be effective in cancer patients carrying *BRCA1* or *BRCA2* mutations and some other ovarian cancers, suggesting synthetic lethality between PARP and other DNA repair pathways (Ström and Helleday, 2012). This supports the potential for PARP inhibition as a chemo-preventative alternative to prophylactic surgery for high risk individuals with *BRCA1* or *BRCA2* mutations (Ström and Helleday, 2012). Hormone-based therapy has also been suggested as a chemo-preventative in such high risk individuals and aromatase inhibitors have completed phase I clinical trials for this purpose (Bozovic-Spasojevic 2012). Ström and Helleday (2012) also postulate increased efficacy of PARP inhibitors in the hypoxic DNA-damaging tumour micro-environment.

A PARP inhibitor, olaparib, showed fewer adverse effects than cytotoxic chemotherapy and anti-tumour activity in phase I trials against *BRCA1* or *BRCA2* deficient familial breast, ovarian, and prostate cancers (Fong *et al.*, 2009) and sporadic ovarian cancer (Fong *et al.*, 2010). AstraZeneca has reported phase II trials showing the treatment is effective in *BRCA1* or *BRCA2* deficient breast (Tutt *et al.*, 2010) and ovarian cancers (Audeh *et al.*, 2010) with a favourable therapeutic window and similar toxicity between carriers of *BRCA1* or *BRCA2* mutations and sporadic cases. As-

traZeneca announced that olaparib has begun phase III trials for breast and ovarian cancers in 2013. Mixed results in phase II trials in ovarian cancer are behind the delays addressed by retrospective analysis of the cohort subgroup with confirmed mutation of *BRCA1* or *BRCA2* genes in the tumour; unsurprisingly these patients, benefit most from the PARP inhibitor treatment and have increased platinum sensitivity in combination treatment. These PARP inhibitors are FDA approved for some cancers McLachlan *et al.* (2016), are effective against germline and sporadic *BRCA1* or *BRCA2* mutations, and are a potential prevention alternative to prophylactic surgery for high risk mutation carriers Ström and Helleday (2012).

This demonstrates the clinical impact of a well characterised system of synthetic lethality with known cancer risk genes. Synthetic lethality has the benefit of being effective against inactivation of tumour suppressor genes by any means, broader than targeting a particular oncogenic mutation (Kaelin, Jr, 2005). The targeted therapy is effective in both sporadic and hereditary *BRCA1* or *BRCA2* deficient tumours acting against an oncogenic molecular aberration across several tissues.

1.2.6 High-throughput Screening for Synthetic Lethality

The function of signalling pathways and combinations of interacting genes are important in cancer research but classical genetics approaches have been limited to non-redundant pathways (Fraser, 2004). The emerging RNAi technologies have vastly expanded the potential for studying genetic redundancy in mammalian experimental models including testing experimentally for synthetic lethality (Fraser, 2004). Identifying synthetic lethality is crucial to study gene function, drug mechanisms, and design novel therapies (Lum *et al.*, 2004). Candidate selection of synthetic lethal gene pairs relevant to cancer has shown some success but is limited because interactions are difficult to predict; they can occur between seemingly unrelated pathways in model organisms (Costanzo *et al.*, 2011). While biologically informed hypotheses have had some success in synthetic lethal discovery (Bitler *et al.*, 2015; Bryant *et al.*, 2005; Farmer *et al.*, 2005), interactions occurring indirectly between distinct pathways would be missed (Boone *et al.*, 2007; Costanzo *et al.*, 2011). Scanning the entire genome for interactions against a clinically relevant gene is an emerging strategy being explored with high-throughput screens (Fece de la Cruz *et al.*, 2015) and computational approaches (Boucher and Jenna, 2013; van Steen, 2012).

Experimental screening for synthetic lethality is an appealing strategy for wider discovery of functional interactions *in vivo* despite many potential sources of error which

must be considered. The synthetic lethal concept has both genetic and pharmacological screening applications to cancer research. Genetic screens, with RNAi to discover the specific genes involved, inform development of targeted therapies with a known mode of action, anticipated mechanisms of resistance, and biomarkers for treatment response. RNAi is a transient knockdown of gene expression more similar to the effect of drugs than complete gene loss and makes comparison to screens in model organisms difficult (Bussey *et al.*, 2006). The RNAi gene knockdown process has inherent toxicity to some cells, potential off-target effects, and issues with a high false positive rate. Therefore, it is important to validate any candidates in a secondary screen and replicate knockdown experiments with a number of independent shRNAs. Alternative gene knockout procedures have also been proposed for synthetic lethal screening including a genome-wide application of the CRISPR/Cas9/sgRNA genome editing technology (Sander and Joung, 2014), episomal gene transfer (Vargas *et al.*, 2004), or RNAi with lentiviral transfection for delivery of shRNA (Telford *et al.*, 2015). Genetic screens have potential for quantitative gene disruption experiments to selectively target overexpressed genes in cancer via synthetic dosage lethality. While powerful for understanding fundamental cellular function, analysis of isogenic cell lines is inherently limited by assuming only a single mutation differs between them despite susceptibility to “genetic drift” and cannot account for diverse genetic backgrounds or tumour heterogeneity (Fece de la Cruz *et al.*, 2015). Genetic screens thus identify targets to develop or repurpose targeted therapies for disease but alone will not directly identify a lead compound to develop for the market or clinical translation.

Chemical screens are immediately applicable to the clinic by directly screening for selective lead compounds with suitable pharmacological properties. However chemical screens lack a known mode of action, may affect many targets, and screen a narrow range of genes with existing drugs. With either approach there are many challenges translating candidates into the clinic such as finding targets relevant to a range of patients, validation of targets, accounting for a range of genetic (and epigenetic) contexts or tumour micro-environment, identifying effective synergistic combinations, enhancers of existing radiation or cytotoxic treatments, avoiding inherent or acquired drug resistance, and developing biomarkers for patients which will respond to synthetic lethal treatment, including integrating these into clinical trials and clinical practice. Identifying specific target genes is an effective way to anticipate such challenges, which can be approached with genetic screens, so we will focus on these and computational alternatives. Screening methods have proven a fruitful area of research, despite being

costly, laborious, and having many different sources of error. These limitations suggest a need for complementary computational approaches to synthetic lethal discovery.

1.2.6.1 Synthetic Lethal Screens

Overexpression of genes is another suitable application for synthetic lethality since over-expressed genes cannot be distinguished from the wild-type by direct sequence specific targeted therapy. Overexpression of oncogenes, such as *EGFR*, *MYC*, and *PIM1*, has been found to drive many cancers. *PIM1* is a candidate for synthetic lethal drug design in lymphomas and prostate cancers, where it interacts with *MYC* to drive cancer growth. van der Meer *et al.* (2014) performed an RNAi screen for synthetic lethality between *PIM1* overexpression and gene knockdown in RWPE prostate cancer cell lines. *PLK1* gene knockdown and drug inhibition was effective as a specific inhibitor of *PIM1* overexpressing prostate cells in cell culture and mouse tumour xenografts. *PLK1* inhibition reduced *MYC* expression in pre-clinical models, consistent with expression in human tumours in which *PIM1* and *PLK1* are co-expressed and correlated with tumour grade. Thus RNAi screening was valuable to identify therapeutic targets and biomarkers for patient response as demonstrated with the finding of *PLK1* as a candidate drug target against prostate cancer progression.

Heredity leiomyomatosis and renal cell carcinoma (HLRCC) is a cancer syndrome of predisposition to benign tumours in the uterus and risk of malignant cancer of the kidney attributed to inherited mutations in fumarate hydratase (*FH*). Boettcher *et al.* (2014) performed an RNAi screen on HEK293T renal cells for synthetic lethality with *FH*. They found enrichment of haem metabolism (consistent with the literature) and adenylate cyclase pathways (consistent with cAMP dysregulation in *FH* mutant cells). Synthetic lethality between *FH* mutation and adenylate cyclases was validated with gene knockdown, drug experiments, and replicated across both HEK293T renal cells and VOK262 cells derived from a HLRCC patient, suggesting new potential treatments against the disease.

Similarly, hereditary diffuse gastric cancer (HDGC) is a cancer syndrome of predisposition to early-onset malignant stomach and breast cancers attributed to mutations in E-cadherin (*CDH1*). Telford *et al.* (2015) performed an RNAi screen on MCF10A breast cells for synthetic lethality with *CDH1*. They found enrichment of G-protein coupled receptors (GPCRs) and cytoskeletal gene functions. The results were consistent with a concurrent drug compound screen with a number of candidates validated by lentiviral shRNA gene knockdown and drug testing including inhibitors of Janus kinase, histone deacetylases, phosphoinositide 3-kinase, aurora kinase, and tyrosine ki-

nases. Therefore the synthetic lethal strategy has potential for clinical impact against HDGC, with particular interest in interventions with low adverse effects for chemo-prevention, including repurposing existing approved drugs for activity against *CDH1* deficient cancers.

RNAi screening for synthetic lethality is also useful for functional genetics to understand drug sensitivity. Aarts *et al.* (2015) screened WiDr colorectal cells for synthetic lethality between *WEE1* inhibitor treatment and an RNAi library of 1206 genes with functions known to be amenable to drug treatment or important in cancer such as kinases, phosphatases, tumour suppressors, and DNA repair (a pathway *WEE1* regulates). Screening identified a number of synthetic lethal candidates including genes involved in cell cycle regulation, DNA replication, repair, homologous recombination, and Fanconi anaemia. Synthetic lethality with cell-cycle and DNA repair genes was consistent with the literature and validation in a panel of breast and colorectal cell lines supported checkpoint kinases, Fanconi anaemia, and homologous recombination as synthetic lethal partners of *WEE1*. These results show that synthetic lethality can be used to improve drug sensitivity as a combination treatment, especially to exploit genomic instability and DNA repair, which are known to be clinically applicable from previous results with *BRCA1* or *BRCA2* genes and PARP inhibitors (Lord *et al.*, 2015). Therefore, *WEE1* inhibitors are an example of treatment which could be repurposed with the synthetic lethal strategy. Similar findings would be valuable to clinicians as a source of biomarkers and novel treatments. While using a panel of cell lines to replicate findings across genetic background is a promising approach to ensure wide clinical application of validated synthetic lethal partners, a computational approach may be more effective as it could account for wider patient variation than scaling up intensive experiments on a wide array of cell lines and could screen beyond limited candidates from an RNAi library.

Chemical genetic screens are also a viable strategy to identify therapeutically relevant synthetic lethal interactions. Bitler *et al.* (2015) investigated *ARID1A* mutations, aberrations in chromatin remodelling known to be common in ovarian cancers, for drug response. Ovarian RMG1 cells were screened for drug response specific to *ARID1A* knockdown cells. They used *ARID1A* gene knockdown for consistent genetic background, with control experiments and 3D cell culture to ensure relevance to drug activity in the tumour micro-environment. Screening a panel of commercially available drugs targeting epigenetic regulators found *ESH2* methyltransferase inhibitors effective and specific against *ARID1A* mutation with validation in a panel of ovarian cell

lines. Synthetic lethality between *ARID1A* and *ESH2* was supported by decreases in H3K27me3 epigenetic marks and markers of apoptosis in response to *ESH2* inhibitors. This was mechanistically supported with differential expression of *PIK3IP1* and association of both synthetic lethal genes with the *PIK3IP1* promoter identifying the PI3K-AKT signalling pathway as disrupted when both genes were inhibited.

This successfully demonstrates the importance of synthetic lethality in epigenetic regulators, identifies a therapeutically relevant synthetic lethal interaction, and shows that chemical genetic screens could model drug response and combination therapy in cancer cells. However this approach is limited to finding synthetic lethal interactions between genes with known similar function, which may not be the most suitable for treatment. Further limiting experiments to genes with existing targeted drugs reduces the number of synthetic lethal interactions detected, assumes the specificity of drugs to a particular target, and many of these drugs are not yet clinically available yet anyway, although they are clinical trials for other diseases or limited to access by patients from a particular countries.

The examples above show that high-throughput screens are an effective approach to discover synthetic lethality in cancer with a wide range of applications. Screens are more comprehensive than hypothesis-driven candidate gene approaches and successfully find known and novel synthetic lethal interactions with potential for rapid clinical application. They have the power to test mode of action of drugs, find unexpected synthetic lethal interactions between pathways, or identify effective treatment strategies without needing a clear mechanism. However, synthetic lethal screens are costly, labour-intensive, error-prone, and biased towards genes with effective RNAi knockdown libraries. Limited genetic background, lethality to wild-type cell during gene knockdown, off-target effects, and difficulty replicating synthetic lethality across different cell lines, tissues, laboratories, or conditions stems from a high false positive rate and a lack of standardised thresholds to identify synthetic lethality in a high-throughput screen. Therefore there is a need for replication, validation, and alternative approaches to identify synthetic lethal candidates. Varied conditions between experimental screens and differences between RNAi and drug screens renders meta-analysis ineffective.

Genome-scale synthetic lethal experiments (across gene pairs) are not feasible, even in model organisms, and they typically focus on specific gene candidates or the partners of a gene of interest (such as importance in health). Therefore a computational approach is more suitable for this task and may further augment experimental screening to replicate screen candidates beyond experimental models.

1.2.7 Computational Prediction of Synthetic Lethality

1.2.7.1 Bioinformatics Approaches to Genetic Interactions

Prediction of gene interaction networks is a feasible alternative to high-throughput screening with biological importance and clinical relevance. There are many existing methods to predict gene networks, as reviewed by van Steen (2012) and Boucher and Jenna (2013) and summarised in Table 1.1. However, many of these methods have limitations including the requirement for existing SGI data, several data inputs, and reliability of gene function annotation. Many of the existing methods also assume conservation of individual interactions between species, which has been found not to hold in yeast studies (Dixon *et al.*, 2008). Tissue specificity is important in gene regulation and gene expression, which are used as predictors of genetic interaction. However, tissue specificity of genetic interactions cannot be explored in yeast studies and has not been considered in many studies of multicellular model organisms, human networks, or cancers. Similarly, investigation into tissue specificity of protein-protein interactions (PPIs), an important predictor of genetic interactions, is difficult given that high-throughput two-hybrid screens occur out of cellular context for multicellular organisms.

Table 1.1: Methods for Predicting Genetic Interactions

Method	Input Data	Species	Source	Tool Offered
Between Pathways Model	PPI, SGI	<i>S. cerevisiae</i>	Kelley and Ideker (2005)	
Within Pathways Model	PPI, SGI	<i>S. cerevisiae</i>	Kelley and Ideker (2005)	
Decision Tree	PPI, expression, phenotype	<i>S. cerevisiae</i>	Wong <i>et al.</i> (2004)	2 Hop
Logistic Regression	SGI, PPI, co-expression, phenotype	<i>C. elegans</i>	Zhong and Sternberg (2006)	Gene Orienteer
Network Sampling	SGI, PPI, GO	<i>S. cerevisiae</i>	Le Meur and Gentleman (2008) Le Meur <i>et al.</i> (2014)	SLGI(R)
Random Walk	GO, PPI, expression	<i>S. cerevisiae</i> <i>C. elegans</i>	Chipman and Singh (2009)	
Shared Function	Co-expression, PPI, text mining, phylogeny	<i>C. elegans</i>	Lee <i>et al.</i> (2010b)	WormNet
Logistic Regression	Co-expression, PPI, phenotype	<i>C. elegans</i>	Lee <i>et al.</i> (2010a)	GI Finder
Jaccard Index	GO, SGI, PPI, phenotype	Eukarya	Hoechndorf <i>et al.</i> (2013)	
Machine Learning			Pandey <i>et al.</i> (2010)	MNMC
Machine Learning Meta-Analysis			Wu <i>et al.</i> (2014)	MetaSL
Flux Variability Analysis				
Flux Balance Analysis	Metabolism	<i>E. coli</i> <i>M. pneumoniae</i>	Güell <i>et al.</i> (2014)	
Network Simulation				

There are a number of existing computational methods for predicting synthetic lethal gene pairs in humans with a specific interest in cancer (as summarised in 1.2). While these demonstrate the power and need for predictions of synthetic lethality in human and cancer contexts, limitations of previous methods could be met with a

Table 1.2: Methods for Predicting Synthetic Lethality in Cancer

Method	Input Data	Source	Tool Offered
Network Centrality	protein-protein interactions	Kranthi <i>et al.</i> (2013)	
Differential Expression	Expression Mutation	Wang and Simon (2013)	
Comparative Genomics	Yeast synthetic gene interactions	Heiskanen and Aittokallio (2012)	
Chemical-Genomics	Homology		
Comparative Genomics	Yeast synthetic gene interactions Homology	Deshpande <i>et al.</i> (2013)	
Machine Learning		Discussed by Babyak (2004) and Lee and Marcotte (2009)	
Differential Expression	Expression	Tiong <i>et al.</i> (2014)	
Literature Database		Li <i>et al.</i> (2014)	Syn-Lethality
Meta-Analysis	Meta-Analysis	Wu <i>et al.</i> (2014)	MetaSL
Pathway Analysis		Zhang <i>et al.</i> (2015)	
Protein Domains	Homology	Kozlov <i>et al.</i> (2015)	
Data-Mining	Expression	Jerby-Arnon <i>et al.</i> (2014)	
Machine Learning	Somatic mutation and DNA CNV siRNA in cell lines	Ryan <i>et al.</i> (2014) Crunkhorn (2014) Lokody (2014)	DAISY (method)
Genome Evolution	Expression	Lu <i>et al.</i> (2013)	
Hypothesis Test	DNA CNV	Lu <i>et al.</i> (2015)	
Machine Learning	Known SL		
Bimodality	Expression DNA CNV Somatic Mutation	Wappett (2014) Wappett <i>et al.</i> (2016)	BImodal Subsetting ExPression (BiSEp)
Directional Chi-Square	Expression (microarray) Somatic mutation	Kelly, S. T., Guilford, P. J., and Black, M. A. Dissertation (Kelly, 2013) and developed here	SLIPT

different approach. Existing computational approaches to synthetic lethal prediction are often difficult to interpret, replicate for new genes, or are reliant on data types not available for a wider range of genes to test.

1.2.7.2 Comparative Genomics

A comparative genomics approach by Deshpande *et al.* (2013) used the results of well characterised high-throughput mutation screens in *S. cerevisiae* as candidates for synthetic lethality in humans (Baryshnikova *et al.*, 2010a; Costanzo *et al.*, 2010, 2011; Tong *et al.*, 2001, 2004). Yeast synthetic lethal partners were compared to human orthologues to find cancer relevant synthetic lethal candidate pairs with direct therapeutic potential. Proposed as a complementary approach to siRNA screens, approximately 24,000 of the 116,000 negative SGIs in yeast (Costanzo *et al.*, 2011) were matched to human orthologues, with over 500 involving a cancer gene (Futreal *et al.*, 2004). Under strict criteria of one-to-one orthologues, large effect size and significant interaction in yeast data ($\epsilon < -0.2$, $p < 0.05$), 1,522 interactions were identified with 70 involving cancer genes. Of the 21 gene interactions tested with pairs of siRNA in IMR1 fibroblast cells, 6 exhibited synthetic lethal effects. The two strongest interactions (*SMARCB1* with

PSMA4 and *ASPSCR1* with *PSMC2*) were successfully validated by protein analysis of human cells and replication with tetrad analysis for yeast orthologues.

Another approach to systematic synthetic lethality discovery specific to human cancer (in contrast to the plethora of yeast synthetic lethality data) was to build a database as done by Li *et al.* (2014). In their relational database, called “Syn-lethality”, they have curated both known experimentally discovered synthetic lethal pairs in humans (113 pairs) from the literature and those predicted from synthetic lethality between orthologous genes in *S. cerevisiae* yeast (1114 pairs). This knowledge-based database is the first dedicated to human cancer synthetic lethal interactions and integrates gene function annotation, pathway and molecular mechanism data with experimental and predicted synthetic lethal gene pairs. This combination of data sources is intended to tackle the trade-off between more conclusive synthetic lethal experiments in yeast and more clinically relevant synthetic lethal experiments in human cancer models, such as RNAi, especially when high-throughput screens are costly and prone to false positives in either system and difficult to replicate across gene backgrounds. This database centralises a wealth of knowledge scattered in the literature including cancer relevant genes (*BRCA1*, *BRCA2*, *PARP1*, *PTEN*, *VHL*, *MYC*, *EGFR*, *MSH2*, *KRAS*, and *TP53*) and is publicly available as a Java App. These included the previously mentioned interactions of *BRCA1* and *BRCA2* with *PARP1* and *TP53* with *WEE1* and *PLK1*. However, the computational methodology was not released, so it is not possible to replicate their results, nor to add to the findings with new datasets, which are limited to 647 human genes. Suggested future directions were promising, such as constructing networks of known synthetic lethality, applying known synthetic lethality to cancer treatment, data mining, replicating the approach for synthetic lethality in model organisms, signalling pathways, and developing a complete global network in human cancer or yeast (both of which are still incomplete with experimental data), some of which has been implemented in “SynLethDB” (Guo *et al.*, 2016).

Machine learning approaches have also been proposed for synthetic lethal discovery (Babyak, 2004; Lee and Marcotte, 2009). Due to concerns that these may be subject to overfitting or noise, Wu *et al.* (2014) developed a meta-analysis method (based on the machine learning methods in Table 1.3) for synthetic lethal gene pairs relevant to developing selective drugs against human cancer, building upon their previous database (Li *et al.*, 2014). They used training data of 10,885 synthetic lethal interactions from yeast experiments of which 7347 occurred between the 5,504 non-essential genes. Their “metaSL” approach utilises genomic, proteomic and annotation data (including GO

Table 1.3: Machine Learning Methods used by Wu *et al.* (2014)

Method	Source	Tool Offered
Random Forest	Breiman (2001)	
Random Forest		
J48 (decision tree)		
Bayes (Log Regression)		
Bayes (Network)	Hall <i>et al.</i> (2009)	WEKA
PART (Rule-based)		
RBF Network		
Bagging / Bootstrap		
Classification via Regression		
Support Vector Machine (Linear)	Vapnik (1995)	
Support Vector Machine (RBF – Gaussian)	Joachims (1999)	
Multi-Network Multi-Class (MNMC)	Pandey <i>et al.</i> (2010)	
MetaSL (Meta-Analysis)	Wu <i>et al.</i> (2014)	MetaSL

terms Ashburner *et al.* (2000), PPI, protein complexes, and biological pathway) with strong statistical performance in yeast data (AUROC of 0.871). The predicted orthologous synthetic lethal partners in human data were not experimentally validated but several would be relevant to cancer such as *EGFR* with *PRKCZ*. They note that computational approaches scale-up across the genome at lower cost than experimental screen and share their most supported interactions online. However, the method is not available for analysis of other genes studied by the cancer research community. While machine learning has great potential as a predictor, the results vary greatly depending on the predictive features selected and it is not clear which threshold should be used to report reliably detected genes. Syn-Lethality (Li *et al.*, 2014) and MetaSL (Wu *et al.*, 2014) demonstrate the value of computational approaches to synthetic lethality but omit many genes of importance in cancer, such as *CDH1*, and there remains a need to enable biological researchers to query such genes in a particular tissue or genetic background.

There is also concern for analyses based on yeast data that many synthetic lethal interactions may not be conserved between species Dixon *et al.* (2009), although interactions between pathways may be more comparable. It is unsurprising that many of the interactions identified were not experimentally validated. There have been many gene duplications in the separate evolutionary histories of humans and yeast which

may lead to differences in genetic redundancy. Yeast are not an ideal human cancer model because they do not have tissue specificity, multicellular gene regulation, or orthologues to a number of known cancer genes such as p53. Although these studies have tried to anticipate these issues with stringent criteria such as requiring one-to-one orthologues, there remains the possibility that changes in gene function may affect whether these are solely redundant such as if functions had coevolved without sequence homology. Many genes will also be excluded since they lack homologues in yeast, the corresponding experimental data, or having paralogues in either species. Thus conservation of yeast interactions is not an ideal strategy and analysis of human data directly for comparison with human experimental data will be the focus of this thesis.

1.2.7.3 Analysis and Modelling of Protein Data

Kranthi *et al.* (2013) took a network approach to discovery of synthetic lethal candidate selection applying the concept to “centrality” to a human PPI network involving interacting partners of known cancer genes. The effect of removing pairs of genes on connectivity of the network was used as a surrogate for viability which is supported by observations that the PPI and synthetic lethal networks are orthogonal in *S. cerevisiae* studies (Tong *et al.*, 2004). They showed that the human cancer protein interaction network (of 1539 proteins and 6471 interactions) exhibits the power law distribution expected of a scale-free synthetic lethal network with high connectivity (average vertex degree of 23.67 and network efficiency of 0.2952). Their top 100 candidate interactions included interactions of the tumour suppressor *TP53* with *BRCA1*, *CDKNA1*, *CDKNA2*, *MET*, and *RB1* which have been detected by prior studies. The gene pairs were often observed to be in the same or a plausible compensatory pathway. Thus the network structure is important in the biological functions of cancers and could be exploited for targeting *TP53* loss of function mutations.

However, their approach was limited to known cancer genes and is not applicable to genes that do not have PPI data. Other nucleotide sequencing data types are more commonly available for cancer studies at a genomic scale. Of further concern is that the results were enriched for p53 synthetic lethal partners which is relevant to many cancer researchers but this genome-wide approach did not detect many other cancer genes due to multiple testing. This enrichment may be due to the known drastic effect of removing p53 itself from the network as a master regulator, cancer driving tumour suppressor gene, and highly connected network “hub”. The focus on cancer genes is useful for translation into therapeutics but does not account for variable genetic backgrounds or effect of protein removal on the whole cellular network.

Focusing on the potential for synthetic lethality to be an effective anti-cancer drug target, Zhang *et al.* (2015) used modelling of signalling pathways to identify synthetic lethal interactions between known drug targets and cancer genes by simulating gene knockdowns. A computational approach was applied to avoid the limitations of experimental RNAi screens such as scale, instability of knockdown, and off-target effects. This ‘hybrid’ method of a data-driven model and known signalling pathways showed potential as a means to predict cell death in single and combination gene knockouts. They used time series protein phosphorylation data (Lee *et al.*, 2012) for 28 signalling proteins and Gene Ontology (GO) (GO) pathways Ashburner *et al.* (2000); Blake *et al.* (2015). This approach successfully detected many known essential genes in the human gene essentiality database, known synthetic lethal partners in the Syn-Lethality database (Li *et al.*, 2014), and predicted novel synthetic lethal gene pairs. The strongest essential genes in single knockdowns were *AKT*, *TP53*, *CHK1*, *S6K1*, and *CYCLIND1*. Pairwise knockdowns identified 252 candidate synthetic lethal interactions including *TP53* with *CHK1*, *S6K1*, *WEE1*, *CYCLIND1*, and *CASP9*; *AKT* with *WEE1*; and *CDK1* with *CYCLIND1*. These novel results contained many *TP53* and *AKT* synthetic lethal partners, genes known to be important in many cancers, however these also have a severe impact on the signalling pathways in their essentiality analysis of single gene disruptions and large phenotypic changes in cancer. This approach is amenable to detect functionally related pathways and protein complexes across the molecular function, cellular component, and biological process annotations provided by GO. The results were consistent with the experimental results in the literature but the novel synthetic lethal interactions have yet to be validated. While the mathematical reasoning and algorithms are given, the code was not released to replicate the findings or apply the methodology beyond the signalling pathways analysed by Zhang *et al.* (2015). While this is an interesting approach, the analysis of this thesis will focus on gene expression and RNAi data which is available to test a wider range of candidate gene pairs.

1.2.7.4 Differential Gene Expression

Differential gene expression has been explored to predict synthetic lethal pairs in cancer which would be widely applicable due to the availability of public gene expression data for a large number of samples and cancer types. Wang and Simon (2013) found differentially expressed genes (by the t-test, adjusted by FDR) between tumours with or without functional p53 mutations in TCGA (McLendon *et al.*, 2008) and Cancer Cell Line Encyclopaedia (CCLE) (Barretina *et al.*, 2012) RNA-Seq gene expression

data as candidate synthetic lethal partner pathways of p53. They identified 2, 8, and 21 candidate synthetic lethal partner genes in 3 microarray datasets from the NCI60 cell lines, 31 partner genes from the CCLE RNA-Seq data (Barretina *et al.*, 2012), and 50 in TCGA RNA-Seq data (Muzny *et al.*, 2012). *PLK1* was replicated across 4 of these analyses and 17 other genes were replicated across 2 analyses (including *MTOR*, *PLK4*, *MAST2*, *MAP3K4*, *AURKA*, *BUB1* and 6 CDK genes) with many playing a role in cell cycle regulation. This was supported by a drug sensitivity experiment on the NCI60 cell lines which found that cells which lacked functional p53 were more sensitive to paclitaxel (which targets *PLK1*, *AURKA*, and *BUB1*). This demonstrated the potential of gene expression as a surrogate for gene function and use of public genomic data to predict synthetic lethal gene pairs in cancer. Wang and Simon (2013) advocated for pre-screening of expression profiles to augment future RNAi screens. However, the analyses were limited to kinase genes and focused on currently druggable genes, lacking wider application of synthetic lethal prediction methodology. This approach may not be feasible or applicable in cancer genes with a lower mutation rate than p53.

Tiong *et al.* (2014) also investigated gene expression as a predictor of synthetic lescal-freethal pairs with colorectal cancer microarrays from a Han Chinese population with a sample size of 70 tumour and 12 normal tissue samples. Simultaneously differentially expression of “tumour dependent” gene pairs (which includes co-expression) between cancer and normal tissue was used to rank 663 candidate synthetic lethal interactions identified in cell line siRNA experiments. Of the top 20 genes, 17 were tested for testing differential expression at the protein level with immunohistochemistry staining and correlation with clinical characteristics, with 11 pairs exhibiting synergistic effects. Some of the predicted synthetic lethal pairs were consistent with the literature (including *TP53* with *S6K1* and partners of *KRAS*, *PTEN*, *BRCA1*, and *BRCA2*) and two novel synthetic lethal interactions (*TP53* with *CSNK1E* and *CTNNB1*) were validated in pre-clinical models. This serves a valuable proof-of-concept for integration of *in silico* approaches to synthetic lethal discovery in cancer demonstrating it’s utility to triage and identify synthetic lethal partners of p53 applicable to colorectal tissues. Although the experimental work was the focus of the paper, these findings show that bioinformatics synthetic lethal candidates can be validated in patient tissue samples (from a non-caucasian population) to find those applicable to colorectal cancers.

1.2.7.5 Data Mining and Machine Learning

Recognising the utility of synthetic lethality to drug inhibition and specificity of anti-cancer treatments, Jerby-Arnon *et al.* (2014) also saw the need for effective prediction

of gene essentiality and synthetic lethality to augment experimental studies of SL. They developed the DAyA mIning SYnthetic lethal identification pipeline (DAISY), a data-driven approach for genome-wide analysis of synthetic lethality in public cancer genomics data from TCGA and CCLE (Barretina *et al.*, 2012). DAISY is intended to predict the candidate synthetic lethal partners of a query gene such as genes recurrently mutated in cancer.

Jerby-Arnon *et al.* (2014) combined a computational approach to triage candidates with a conventional RNAi screen to validate synthetic lethal partners. They screened a selection of computationally predicted candidates and randomly selected genes with RNAi against *VHL* loss of function mutation in RCC4 renal cell lines. The computational method had a high AUROC of 0.779 and predictions were enriched 4× for validated RNAi hits over randomly selected genes. This approach detected known synthetic lethal pairs such as *BRCA1* or *BRCA2* genes with *PARP1* and *MSH2* with *DHFR*. The synthetic lethal candidates identified with both RNAi screening and computational prediction formed an extensive network of 2077 genes with 2816 synthetic lethal interactions and similar network of 3158 genes with 3635 synthetic dosage lethal interactions (for synthetic lethality with over-expression). Each network was scale-free as expected of a biological network and was enriched for known cancer genes, essential genes in mice, and could be harnessed for predicting prognosis and drug response. While demonstrating the feasibility of combining experimental and computational approaches to synthetic lethality in cancer, there remain challenges in predicting synthetic lethal genes, novel drug targets, and translation into the clinic.

The DAISY methodology (Jerby-Arnon *et al.*, 2014) compares the results of analysis of several data types to predict synthetic lethality, namely: DNA copy number and somatic mutation for TCGA patient samples and CCLE cell lines. The cell lines were also analysed with gene expression and gene essentiality (shRNA screening) profiles. Genes were classed as inactivated by copy number deletion, somatic loss of function mutation, or low expression and tested for synthetic lethal gene partners which are either essential in screens or not deleted with copy number variants. Co-expression is also used for synthetic lethality prediction based on studies in yeast (Costanzo *et al.*, 2010; Kelley and Ideker, 2005). Copy number, gene expression and, essentiality analyses are stringently compared by adjusting each for multiple tests with Bonferroni correction and only taking hits which occur in all analyses. This methodology was also adapted for synthetic dosage lethality by testing for partner genes where genes are overactive with high copy number or expression. As discussed above, the predictions performed

well and an RNAi screen for the example of *VHL* in renal cancer validated predicted synthetic lethal partners of *VHL* demonstrating the feasibility of combining approaches to synthetic lethal discovery in cancer and using computational predictions to enable more efficient high-throughput screening. DAISY performs well statistically with a AUROC of 0.779 on a set of gene pairs with experimental screen data, although co-expression and shRNA functional examination contributes much less of this than the mutation and copy number analysis (AUROC 0.683 alone). However, this methodology is very stringent, missing potentially valuable synthetic lethal candidates, may not be applicable to genes of interest to other groups and the software for the procedure has not been publicly released for replication.

Although the DAISY procedure performs well and has been well received by the scientific community (Crunkhorn, 2014; Lokody, 2014; Ryan *et al.*, 2014), showing a need for such methodology, there is no indication of adoption of the methodology in the community yet. The co-expression analysis may not be the most effective way to test gene expression for directional synthetic lethal interactions (where inverse correlation would be expected). In the interests of a large sample size, tissue types were not tested separately despite tissue-specific synthetic lethality being likely since gene function (and by extension expression, isoforms, and clinical characteristics) in cancers may often be tissue-dependent. Some data forms and analyses used, such as gene essentiality, may not be available for all cancers, genes, or tissues, and may not be reproduced.

Lu *et al.* (2015) critique the assumption of co-expression in the DAISY methodology and propose an alternative computational prediction of synthetic lethality based on machine learning methods and a “cancer genome evolution” hypothesis. Using DNA copy number and gene expression data from TCGA patient samples, a cancer genome evolution model assumes that synthetic lethal gene pairs behave in two distinct ways in response to an inactive synthetic lethal partner gene, either a “compensation” pattern where the other synthetic lethal partner is overactive or a “co-loss underrepresentation” pattern where the other synthetic lethal partner is less likely to be lost, since loss of both genes would cause death of the cancer cell. During the genome evolution of cancers, the cell becomes addicted to the remaining synthetic lethal partner due to induced gene essentiality. These patterns would explain why DAISY detects only a small number of synthetic lethal pairs, compared to the large number expected based on model organism studies (Boone *et al.*, 2007), and the disparity between screening and computationally predicted synthetic lethal candidates due to testing different classes of synthetic lethal gene pairs.

Lu *et al.* (2015) compared a genome-wide computational model of genome evolution and gene expression patterns to the experimental data of Vizeacoumar *et al.* (2013) and Laufer *et al.* (2013). This simpler model performing well with an AUROC of 0.751 but was less than DAISY, although it did not rely on data from cell lines which may not represent patient disease. They predict a larger comprehensive list of 591,000 human synthetic lethal partners with a probability score threshold of 0.81, giving a precision of 67% and 14 \times enrichment of synthetic lethal true positives compared to randomly selected gene pairs. Discovery of such a vast number of cancer-relevant synthetic lethal interactions in humans would not be feasible experimentally and is a valuable resource for research and clinical applications. These predictions are not limited by assuming co-expression of synthetic lethal partners or evolutionary conservation with model organisms enabling wider synthetic lethal discovery. However, there remains a lack of basis for an expectation of how many synthetic lethal partners a particular gene will have, how many pairs there are in the human genome, and whether pathways or correlation structure would influence predicted synthetic lethal partners.

Large scale, computational approaches have yet to determine whether synthetic lethal interactions are tissue-specific since Lu *et al.* (2015) used pan-cancer data for 14136 patients with 31 cancer types. Experimental data used for comparison was a small training dataset specific to colorectal cancer, and based on screens for other phenotypes, which may limit performance of the model or application to other cancers. Proposed expansion of the computational approach to mutation, microRNA, or epigenetic modulation of gene function and tumour micro-environment or heterogeneity suggests that synthetic lethal discovery could be widely applied to the current challenges in cancer genomics. This approach was also based on machine learning methodology and not supported by a software release for the community to develop, contribute to, or reproduce beyond the gene pairs given in the supplementary results.

1.2.7.6 Bimodality

Wappett *et al.* (2016) demonstrate a multi-omic approach to identification of synthetic lethality in cancer with a strategy to detect bimodal patterns in molecular profiles. They released this solution as the BiSEp R package Wappett (2014) which aims to detect subtle bimodal and non-normal patterns in expression data. Since loss of gene function is not consistently genetic, Wappett *et al.* (2016) advocate the use of gene expression (loss of mRNA) and deletion (loss of copy number) data in addition to mutation. The BiSEp procedure was demonstrated on an analysis of 881 cell lines from CCLE (Barretina *et al.*, 2012), 442 cell lines from Catalogue Of Somatic Mutations In

Cancer (COSMIC) (Forbes *et al.*, 2015), and RSEM normalised RNA-Seq data for 178 TCGA lung patient samples (Collisson *et al.*, 2014). BiSEp was demonstrated to have significant enrichment of validated tumour suppressor, synthetic lethal gene pairs (detecting 76 experimentally supported gene pairs) and was improved (detecting 420) with expression data rather than relying on detecting loss of gene function by mutation or deletion. They identified interactions with genes relevant to cancer with support in experimental screens including *ERCC4* with *XRCC1*, *BRCA1* with *PARP3*, and *SMARCA1* with *SMARCA4*.

Wappett *et al.* (2016) demonstrated that analysis of genomics data, particularly expression data, is relevant to augment the identification of synthetic lethal interactions with screening experiments. They further show that this is applicable in both genetically homogeneous cell lines and heterogeneous cell population from patient samples. This approach is limited however to genes which exhibit bimodal expression patterns which do not commonly occur, particularly in normalised gene expression data, and other approaches may need to be considered for gene such as *CDH1* which were not identified by BiSEp.

1.2.7.7 Rationale for Further Development

Many of the approaches discussed here aimed to identify the strongest synthetic lethal pairs across the yeast or human genomes (Deshpande *et al.*, 2013; Lu *et al.*, 2015; Wappett *et al.*, 2016; Wu *et al.*, 2014), which may not be an ideal strategy to identify interactions in particular functions or relevance to particular cancers. These demonstrate a need for computational approaches to prioritise candidate gene pairs for validation but this thesis will focus on the interactions with *CDH1* with particular importance in breast and stomach cancers, although these partners may be applicable in other cancers. As such, this thesis presents a query-based method, amenable to identification of candidate partners for a selected gene of functional or translational importance such as *CDH1*.

1.3 E-cadherin as a Synthetic Lethal Target

E-cadherin is a transmembrane protein (encoded by *CDH1*) with several characterised functions in the cytoskeleton and cell-to-cell signalling. Here we outline the key known functions of E-cadherin and its importance in cancer biology. *CDH1* is a tumour suppressor gene, with loss of function occurring in both familial (germline mutations) and sporadic (somatic mutations) cancers. As such, *CDH1* inactivation is a prime example of a genetic event that could be targeted by synthetic lethality for anti-cancer

treatments. Most notably this includes patients at risk of developing hereditary breast and stomach cancers for which conventional surgical or cytotoxic chemotherapy is not ideal (due to impact of quality of life) and who have a known genetic aberration in their familial syndromic cancers. Effective treatments against *CDH1* inactivation would also benefit patients with sporadic diffuse gastric cancers since they often present with symptoms at a late stage.

1.3.1 The *CDH1* gene and it's Biological Functions

The tumour suppressor gene *CDH1* is implicated in hereditary and sporadic lobular breast cancers (Berx *et al.*, 1996; Berx and van Roy, 2009; De Leeuw *et al.*, 1997; Masciari *et al.*, 2007; Semb and Christofori, 1998; Vos *et al.*, 1997). The *CDH1* gene encodes the E-cadherin protein and is normally expressed in epithelial tissues, where it has also been identified as an invasion suppressor and loss of *CDH1* function has been implicated in breast cancer progression and metastasis (Becker *et al.*, 1994; Berx *et al.*, 1995; Christofori and Semb, 1999).

1.3.1.1 Cytoskeleton

The primary function of *CDH1* is cell-cell adhesion forming the adherens junction, maintaining the cytoskeleton and mediating molecular signals between cells. The function of the adherens complex is particularly important for cell structure and regulation because it interacts with cytoskeletal actins and microtubules. The cytoskeletal role of E-cadherin maintains healthy cellular viability and growth in epithelial tissues including cellular polarity. E-cadherin is not essential to cellular viability but loss in epithelial cells does lead to defects in cytoskeletal structure and proliferation. In addition to a central role in the adherens complex, E-cadherin is involved in many other cellular functions and thus *CDH1* is regarded as a highly pleiotropic gene.

1.3.1.2 Extracellular and Tumour Micro-Environment

As a transmembrane signalling protein E-cadherin also interacts with the extracellular environment and other cells, most notably forming tight junctions between cells. These junctions serve to both regulate movement of ion signals between cells and separate membrane proteins on the apical and basal surfaces of a cell, maintaining cell polarity. Thus E-cadherin is an important regulator of epithelial tissues by intercellular communication. It also has important roles in the extracellular matrix, including fibrin clot formation. The role of intercellular interactions and the tissue micro-environment are important themes in cancer research, being a potential mechanism for cancer progres-

sion and malignancy in addition to its potential for specifically targeting tumour cells.

1.3.1.3 Cell-Cell Adhesion and Signalling

The signals mediated by tight junctions are also passed on to intracellular signalling pathways and thus E-cadherin also has a role in maintaining cellular function and growth. One such example is the regulation of β -catenin which interacts with both the actin cytoskeleton and acts as a transcription factor via the WNT pathway. Similarly, the Hippo and PI3K/AKT pathways are implicated in being mediated by E-cadherin, having roles in promoting cell survival, proliferation, and repressing apoptosis. E-cadherin shares several downstream pathways with signalling pathways such as integrins and thus indirectly interacts with them, particularly since feedback loops may occur in such pathways. Conversely, the multifaceted roles of E-cadherin have been shown with differing overexpression in ovarian cells promoting tumour growth, while it maintains healthy cellular functions in other cells.

1.3.2 *CDH1* as a Tumour (and Invasion) Suppressor

E-cadherin has key roles in maintaining cellular structure and regulating growth, consistent with *CDH1* being a tumour suppressor gene. Loss of *CDH1* in epithelial tissues leads to disrupted cell polarity, differentiation, and migration. E-cadherin loss has been identified as a recurrent driver tumour suppressor mutation in sporadic cancers of many tissues including breast, stomach, lung, colon, and pancreas tissue.

1.3.2.1 Breast Cancers and Invasion

E-cadherin loss in breast cancers has been shown to cause increased proliferation, lymph node invasion, and metastasis with poor cell-cell contact. Thus *CDH1* gene has also been implicated as an invasion suppressor, with a key role in the epithelial-mesenchymal transition (EMT), an established mechanism of cancer progression (Hanahan and Weinberg, 2011). The epithelial-mesenchymal transition is important during development and wound healing but such changes in cellular differentiation also occur in cancers. If *CDH1* is inactivated by mutation or DNA methylation (Berx *et al.*, 1996; Guilford, 1999; Machado *et al.*, 2001), it is likely that EMT will drive growth of E-cadherin deficient cancers (Berx and van Roy, 2009; Graziano *et al.*, 2003; Polyak and Weinberg, 2009). While loss of E-cadherin is not sufficient to cause EMT or tumourigenesis, it is an important step in this mechanism of tumour progression and a potential therapeutic

intervention may therefore also impede cancer progression and have activity against advanced stage cancers.

1.3.3 Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer

CDH1 loss of function mutations also causes familial cancers, including diffuse gastric cancer and lobular breast cancer (Graziano *et al.*, 2003; Guilford *et al.*, 2010, 1999; Oliveira *et al.*, 2009). Individuals carrying a null mutation in *CDH1* have a syndromic predisposition to early-onset these cancers, known as hereditary diffuse gastric cancer (HDGC) (Guilford *et al.*, 1998). Due to the loss of an allele, these individuals are prone to carcinogenic lesion in the breast and stomach when the other allele is inactivated, occurring much more frequently and thus younger than in individuals without a second functional allele of *CDH1*. The loss of the second allele is most often hypermethylation suppressing expression rather than mutation, although loss of heterozygosity may also occur. Therefore HDGC is an autosomal dominant cancer syndrome with incomplete penetrance. The “lifetime” (until age 80 years) risk for mutation carriers of diffuse gastric cancer is 70% in males and 56% in females. In addition, the lifetime risk of lobular breast cancer is 42% in female mutation carriers.

HDGC affects less than 1 in a million people globally (Ferlay *et al.*, 2015) and less than 1% of gastric cancers. However, HDGC is documented to affect several hundred families globally. E-cadherin mutations in the germline is implicated in 1-3% of gastric cancers presenting with a family history, varying between high and low incidence populations. E-cadherin is also mutated in 13% of sporadic gastric cancers.

While diagnostic testing for *CDH1* genotype has enabled more effective management of HDGC and improved patient outcomes, there are still limited options for clinical interventions (Guilford *et al.*, 2010). Individuals with a family history of HDGC are recommended to be tested for *CDH1* mutations in late adolescence and are offered prophylactic stomach surgery before the risk of developing cancers increases with age. Another option is annual endoscopic screening to diagnose early stage stomach cancers with surgical intervention once they are detected (Oliveira *et al.*, 2013). However, these early stage cancers are difficult to detect and may be missed in regular screening. Thus patients carrying *CDH1* mutations either have surgical interventions with a significant impact on quality of life and risk of complications or remain at risk of developing advanced stage stomach cancers. Due to the lower mortality rate due to stomach cancers,

there is increasing concerns among these HDGC families on the elevated risk of lobular breast cancers for women later in life.

The current clinical management of HDGC still has significant risks for patients and therefore a greater understanding of the molecular and cellular function of *CDH1* is important for its role in these cancers. Such studies may lead to alternative treatment strategies such as pharmacological treatments with specificity against *CDH1* null cells, once they lose the second allele. While a loss of gene function cannot be targeted directly, designing a treatment with specificity against *CDH1* may also have activity in sporadic cancers in a range of epithelial cancers. Thus an effective treatment against *CDH1* mutant cancers would potentially have significant therapeutic and preventative applications in a large number of patients.

1.3.4 Models of *CDH1* loss in cell lines

Previous work our research group has published used a model of homozygous *CDH1*^{-/-} null mutation in non-malignant MCF10A breast cells to show that loss of *CDH1* alone was not sufficient to induce EMT with compensatory changes in the expression of other cell adhesion genes occurring (Chen *et al.*, 2014). However, *CDH1* deficient cells did manifest changes in morphology, migration, and weaker cell adhesion (Chen *et al.*, 2014).

This *CDH1*^{-/-} MCF10A model has been used in a genome-wide screen of 18,120 genes using short interfering ribonucleic acid (siRNA) and a complementary drug screen using 4,057 compounds to identify synthetic lethal partners to E-cadherin (Telford *et al.*, 2015). One of the strongest candidate pathways identified by Telford *et al.* (2015) were the GPCR signalling cascades, which were highly enriched by GO analysis of the candidate synthetic lethal partners the primary siRNA screen. This was supported by validation with Pertussis toxin, known to target G_{αi} signalling (Clark, 2004), as were various candidate cytoskeletal pathways by inhibition of Janus kinase (JAK) and aurora kinase. The drug screen also produced candidates in histone deacetylase (HDAC) and phosphoinositide 3-kinase (PI3K) which were supported by validation and time course experiments.

1.4 Summary and Research Direction of Thesis

Genomics technologies and the data available from them have immense potential for understanding of genetics and improving healthcare, including identification of genes altered in cancer for molecular diagnosis, prognostic biomarkers, and therapeutic targets.

This has been demonstrated with the identification of cancer genes in many cancers, distinguishing tumour subtypes by expression profiles, and the development of targeted therapies against oncogenes (such as *BRAF*) and tumour suppressors (such as *BRCA1*). Synthetic lethality is an important genetic interaction to study fundamental cellular functions and exploit them for biomarkers and cancer treatment. They present a means to target loss of function mutations and genetic dysregulation in tumour suppressor genes by identifying interacting partners with redundant or compensating molecular functions.

CDH1 (encoding E-cadherin) is an example of a tumour suppressor gene implicated in sporadic breast and stomach cancers. Germline mutations in *CDH1* are also found in many patients with familial early onset cancers (HDGC). Discovery of synthetic lethal partners would contribute to an understanding on the molecular mechanisms driving the growth of *CDH1* deficient tumours and identification of potential therapeutic targets or chemopreventative agents for management of HDGC. The clinical potential of the synthetic lethal approach has been demonstrated with the application of olaparib against *BRCA1* and *BRCA2* mutations Lord *et al.* (2015) but there remains the need to systematically identify synthetic lethal partner genes for other tumour suppressors such as *CDH1*. A synthetic lethal screen has been conducted on breast cell lines Telford *et al.* (2015) but computational approaches to identification of synthetic lethal partners of *CDH1* remains to be done.

While there are a wide range of experimental and computational approaches to synthetic lethal discovery, many are limited to particular applications, prone to false positives, inconsistent across independent approaches, or enriched for particular genes of interest. Therefore synthetic lethal interactions are difficult to replicate or apply in the clinic. Computational approaches to synthetic lethality are not widely adopted by the cancer research community and experimental approaches cannot be combined to study synthetic lethality at a genome-wide scale. However, these show interest in synthetic lethal discovery in the community and the need for robust predictions of synthetic lethal interactions in cancer and human tissues.

Effective screening, prediction, and analysis of synthetic lethal interactions are a crucial part of developing next generation anti-cancer strategies. Therefore, we propose developing a computational statistical procedure to identify synthetic lethal interactions and construct gene networks. This will enable the development of personalised medicine targeted to particular molecular aberrations. Genetic tests and genomics have the potential to revolutionise cancer screening, diagnosis, and prognostics; tar-

geted therapeutics, similarly, have applications in prevention and therapy of sporadic or hereditary cancers with known molecular properties.

To address the concerns raised by recent computational approaches to synthetic lethal discovery in cancer (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015; Wappett *et al.*, 2016), I present similar analysis using solely gene expression data which is widely available for a large number of samples in many different cancers. This uses a statistical methodology the Synthetic Lethal Interaction Prediction Tool (SLIPT) developed for this purpose. To further determine the limitations and implications of synthetic lethal predictions, modelling and simulation was performed upon the statistical behaviour of synthetic lethal gene pairs in genomics data. Comparison of synthetic lethal gene candidates from public data analysis and experimental candidates, pathway analysis, and networks structure will also be presented to investigate the relationships between synthetic lethal candidates. Release of R codes used for simulation, prediction, and analysis will enable adoption of the methodology in the cancer research community and comparison to existing methods. Therefore my thesis aims to develop such predictions for synthetic lethal partner genes with a focus on the example of E-cadherin to compare to the findings of Telford *et al.* (2015), develop of network approaches for pathway structure, and simulate gene expression on pathway structure with the bioinformatics and computational biology investigations.

1.4.1 Thesis Aims

Understanding synthetic lethality is important in cancers and has shown an impact clinical practice and patient outcomes for certain genes already. Thus this thesis aims to identify synthetic lethal gene pairs using public gene expression data. Accordingly, chapter 3 aims to develop the methods to do so including a synthetic lethal detection methodology (SLIPT) which has been released as an R software package. This chapter also documents the simulation and network analysis procedures developed to support the use of SLIPT and perform analyses in throughout this thesis.

This thesis also aimed to demonstrate SLIPT methodology for analysis of RNA-Seq gene expression data. Chapter 4 aimed to do so by performing an analysis to identify candidate synthetic lethal gene partners of *CDH1* in public breast and stomach cancer data (Bass *et al.*, 2014; TCGA, 2012). Chapter 4 also aimed to demonstrate the biological relevance of these candidate synthetic lethal partners by identifying synthetic lethal pathways and comparing them with the results of an experimental siRNA screen (Telford *et al.*, 2015).

Pathway analysis was extended with graph structures in chapter 5 which aimed to assess the importance of synthetic lethal genes within pathway structures. Chapter 5 also aimed to use pathway structure to identify directional relationships between SLIPT and siRNA synthetic lethal candidates and explore the disparity between them. The SLIPT methodology was also supported by simulations investigations in chapters 3 and 6 which aimed to evaluate the SLIPT method on detecting known synthetic lethal genes in simulated data. Graph structures were also used in chapter 6 which also aimed to determine the effect of pathway structures of synthetic lethal detection with SLIPT in simulated data and ascertain that the simulation results were comparable to expression data with contains complex correlation structures within biological pathways.

Chapter 2

Methods and Resources

In this Chapter, I will outline the various existing resources and methods utilised throughout this project. This includes public data repositories, stable and development releases of software packages (mostly for the R programming environment), and implementation of bioinformatics methods and statistical concepts with Shell or R scripts developed for this purpose. Methods and packages developed specifically for this project will be covered in more detail along with preliminary data to demonstrate and support their use in Chapter 3.

2.1 Bioinformatics Resources for Genomics Research

2.1.1 Public Data and Software Packages

Various bioinformatics resources, such as databases and methods, have become integral parts of genetics and genomics research. Reference genomes, genotyped variants, gene expression, and epigenetics profiles are among the most commonly used resources. Gene expression data in particular is widely available from many microarray and RNA-Seq studies, from repositories such as Gene Expression Omnibus (GEO) (Clough and Barrett, 2016), caArray (Heiskanen *et al.*, 2014), and ArrayExpress (Rustici *et al.*, 2013). Such profiles are excellent resources to examine the changes of gene expression occurring in cancers and the variation between samples. These microarray initiatives have set a precedent for data sharing, data mining, and the wider benefits of publicly available data for enabling the scientific community to further utilise the data rather than a single research group or consortium (Rung and Brazma, 2013). The practice of integrating findings from publicly available genomics data with the research questions and experimental results of individual research groups has carried over into RNA-Seq

datasets including the large-scale cancer genomics projects (Zhang *et al.*, 2011). This thesis is one such example of an investigation enabled by this wider movement and tools developed in various disciplines to generate, process, and disseminate genomic-scale data.

Along with databases, it is also becoming common practice for bioinformatics researchers to release their code as open-source or provide a software package to enable replication of the findings or further applications of the methods (Stajich and Lapp, 2006). This is part of a wider movement in software and data analysis with many tools to facilitate such work being released for use in Linux or the R programming environment (R Core Team, 2016). In addition to the R packages hosted on CRAN (CRAN, 2017), the Bioconductor repositories (Gentleman *et al.*, 2004) also contain many packages specifically for applications in bioinformatics, and the GitHub site hosts many packages in various stages of development and early release. Packages from these various sources have been used throughout this project and cited where-ever possible. Several R packages have been developed during this thesis project and either publicly released on GitHub or prepared to accompany a publication.

2.1.1.1 Cancer Genome Atlas Data

Molecular profile data from normal and tumour samples was downloaded from publicly available sources, using the TCGA (TCGA, 2012) and the International Cancer Genome Consortium (ICGC) web portals (Zhang *et al.*, 2011). These include gene expression (RNA-Seq), somatic mutations, and anonymous clinical data. These versions downloaded were on the 6th of August 2015 (Release 19) and the 2nd of May 2016 (Release 20) for breast and stomach cancer respectively via the ICGC data portal (<https://dcc.icgc.org/>).

Performing a genomic alignment remains a challenge in bioinformatics and methods to do so may yet be improved (Chen and Tompa, 2010). However, the statistical and biological aspects of bioinformatics are the focus of this thesis, comparing alignment methods is outside the scope of these investigations. The TCGA project (TCGA, 2012) used widely adopted tools: “Bowtie” for alignment (Langmead *et al.*, 2009), “mapsplice” to detect splice sites (Wang *et al.*, 2010), and the Reads Per Kilobase per Million mapped reads (RPKM) approach to qualify reads per transcript as a measure of gene expression (Mortazavi *et al.*, 2008). These are widely acceptable tools for processing RNA-Seq data which were used to produce the raw counts of mapped reads (tier 1) and normalised expression data (tier 3) publicly available from TCGA.

Raw count and RSEM normalised TCGA expression data from Illumina RNA-Seq protocols were available from 1,177 breast samples (113 normal, 1,057 primary tumour, and 7 metastases) for 20,501 genes. TCGA breast somatic mutation data for 981 samples (976 primary tumours and 5 metastases) across 25,836 genes were available including 969 samples (964 primary tumours and 5 metastases) with corresponding RNA-Seq expression data and 19,166 genes mapped from Ensembl identifiers to gene symbols, of which 16,156 had corresponding gene expression information. Unless otherwise stated, the raw counts were used for further processing rather than the RSEM normalised data (provided by TCGA tier 3). For the purposes of this analysis somatic mutations were reported if they were detected to non-synonymous substitutions, frameshifts, or truncations (by premature stop codons) which would likely disrupt the wild-type gene function. Normalised protein expression was used (as provided by TCGA tier 3), generated from reverse phase protein arrays (RPPA) for 142 antibodies targeting 115 genes for 298 TCGA breast samples.

Raw count TCGA expression data (TCGA tier 1) from Illumina RNA-Seq was also available for 450 stomach samples (35 normal, 415 primary tumour) for 20,501 genes. TCGA stomach mutation data was also available for 289 samples across 25807 genes, corresponding to 19436 genes with expression data. Normalised protein expression (RPPA) data was also sourced (from TCGA tier 3) for 201 antibodies targeting 158 genes for 357 stomach samples.

2.1.1.2 Reactome and Annotation Data

Unless otherwise specified, pathway analysis was performed for human pathway annotation from the Reactome database (version 52) with pathway gene sets derived from the `reactome.db` R package. Entrez identifiers were mapped to gene symbols or aliases to match to TCGA expression and mutation data using the `org.Hs.eg.db` R package. Further pathway analysis used breast cancer gene signatures from Gatza and colleagues (Gatza *et al.*, 2011; Gatza *et al.*, 2014). These gene symbols were matched to the relevant dataset and used to construct a matrix of category membership using the `safe` R package (Barry, 2016).

2.2 Data Handling

2.2.1 Normalisation

Apart from the Prediction Analysis of Microarray 50 (PAM50) subtyping procedure (Parker *et al.*, 2009), which required RSEM normalised data (J.S. Parker personal

communication), the analysis of the RNA-Seq data presented here was based on raw read count data. Raw read counts were log-scaled; samples removed for consistency (based on a Euclidean distance correlation matrix as described in Section 2.2.2); and the final dataset was TMM normalised (Robinson and Oshlack, 2010) then processed using the `voom` function (Law *et al.*, 2014) in the `limma` R package (Ritchie *et al.*, 2015). Protein expression data generated from RPPA was normalised to dilution curves using the `SuperCurve` R package (Ju *et al.*, 2015; Neeley *et al.*, 2009).

2.2.2 Sample Triage

The TCGA breast RNA-Seq data were assessed for batch effects using a correlation matrix of the log-transformed raw counts for which a heatmap (Euclidean distance, complete linkage) is shown in Figure A.2. While no major batch effects were detectable between the samples, 9 samples were excluded due to poor correlation with the remaining samples, as detailed in Table 2.1. These samples showed unusual density plots compared to the rest of the dataset, and exhibited low mean read count in Figures 2.1 and 2.2. A heatmap showing key clinical properties of these excluded samples and their correlation with the remainder of the samples is shown in Figure A.1, and a full correlation heatmap (Figure A.2) shows these samples as relatively poorly correlated outliers in the bottom rows and left columns. In addition to the clustering analysis (in Appendix A.1), replicate tumour samples were also examined for sample quality in Appendix A.2. After removal of these samples, the TCGA dataset used for analysis consisted of the remaining 1168 samples (from 1040 patients): 1049 tumour samples, 112 normal tissue for matched samples, and 7 metastases.

Table 2.1: Excluded Samples by Batch and Clinical Characteristics.

Tissue Source	Type	Batch	Plate	Patient	Samples	p53	Subtype	Treatment (History)	Clinical Subtypes (Stage)
A7 Christiana	Tumour	47	A227	A0DB	1 of 3	NA	Luminal A	Mastectomy	(no) ER+ Ductal (2)
A7 Christiana	Tumour	96	A220	A13D	1 of 3	Wildtype	Luminal A	Mastectomy	(no) ER+ Ductal (2)
A7 Christiana	Tumour	96	A227	A13E	1 of 3	NA	Basal	Lumpectomy	(no) ER- Ductal (2)
A7 Christiana	Tumour	142	A277	A26E	1 of 3	NA	Basal	Lumpectomy	(no) ER+ Ductal (2)
A7 Christiana	Tumour	47	A277	A0DC	1 of 2	NA	Luminal A	Mastectomy	(yes) ER+ Lobular (3)
A7 Christiana	Tumour	142	A220	A26I	1 of 2	Mutant	Basal	Lumpectomy	(yes) ER- Ductal (2)
AC Intl Genomics	Tumour	177	A18M	A2QH	2 of 2	Mutant	Basal	Radical Mastectomy	(no) ER- Metaplastic (2)
AC Intl Genomics	Tumour	177	A220	A2QH	2 of 2	Mutant	Basal	Radical Mastectomy	(no) ER- Metaplastic (2)
GI ABS IUPUI	Normal	177	A16F	A2C8	1 of 1	NA	Luminal A	Radical Mastectomy and Neoadjuvant	(no) ER+ Ductal (2)

Similarly, a correlation matrix of log-transformed raw counts was used to evaluate sample quality for TCGA stomach RNASeq. A tumour sample (patient 4294) was removed due to similar quality concerns leaving a final dataset for 449 samples (from 417 patients): 414 tumour samples and 35 normal tissue samples.

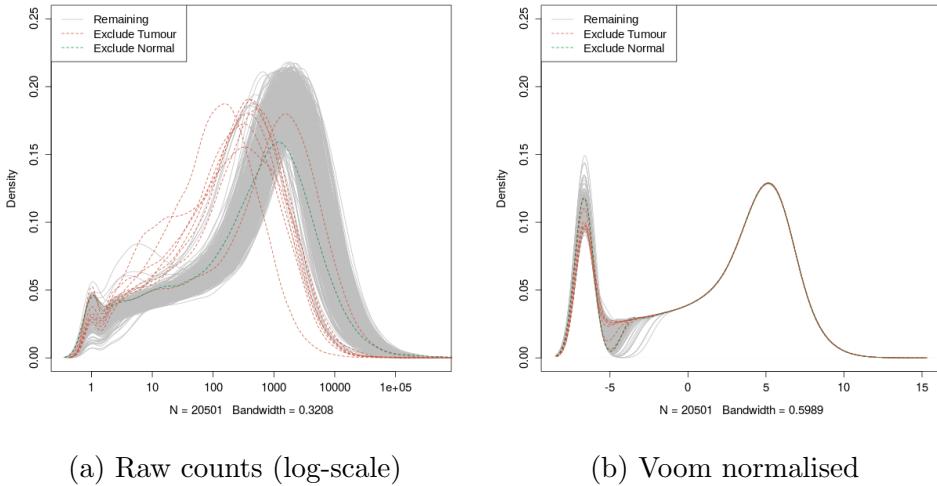


Figure 2.1: **Read count density.** Sample density plots of raw counts on log-scale and voom normalised showing samples removed due to quality concerns.

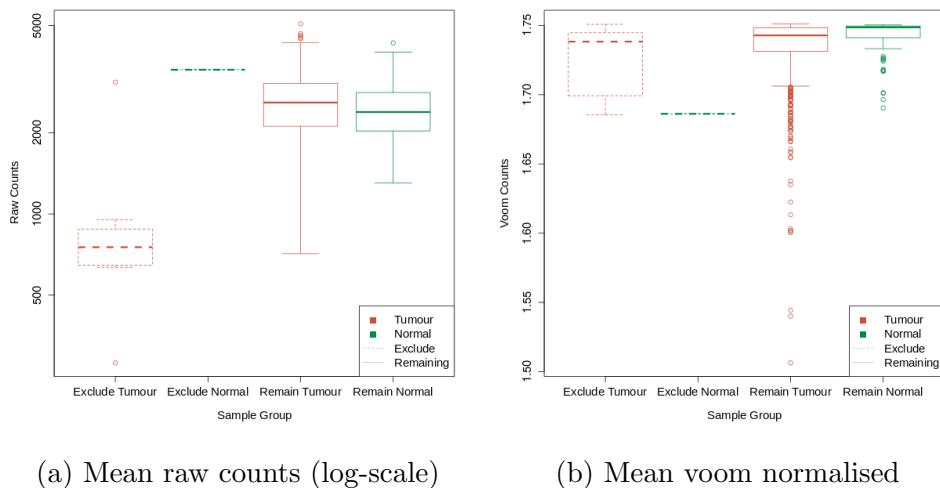


Figure 2.2: **Read count sample mean.** Boxplots of sample means for raw counts on log-scale and voom normalised show removed tumour samples with low mean read count.

2.2.3 Metagenes and the Singular Value Decomposition

A “metagene” offers a consistent signal of pathway (expression) activation or inactivation by dimension reduction of a matrix, avoiding negatively correlated genes averaging out the signal of a mean-based centroid (Huang *et al.*, 2003). Constructing these pathway metagenes used gene sets for Reactome and Gatzka signatures (Gatzka *et al.*, 2011, 2014) as specified above (see Section 2.1.1.2). The singular-value decomposition was performed ($X = U^T DV$ where X is the data matrix of the gene set with genes \times samples) and the leading eigenvector (first column of V) corresponding to the largest singular value was used as a metagene for the pathway gene set. To ensure consistent directionality of metagene signals, the median of the gene set in each sample was calculated and correlated against the metagene with the (arbitrary) metagene sign adjusted as needed to conform with the majority of the gene set (i.e., positive correlation between metagene and the median-based centroid). To ensure that genes and pathways were weighted equally, metagenes were derived from a z-transformed dataset of gene expression and samples were scaled (by fractional ranking) for each metagene so that they were comparable on a [0, 1] scale.

2.2.3.1 Candidate Triage and Integration with Screen Data

Candidate triage in combination with the experimental data was intended to integrate findings of the SLIPT analysis with an ongoing experiment project (Chen *et al.*, 2014; Telford *et al.*, 2015). The first procedure to compare the SLIPT gene candidates for *CDH1* with an siRNA experimental screen (Telford *et al.*, 2015) was a direct comparison of the overlapping candidates, presented in a Venn diagram and tested with the χ^2 test. Since these candidates modestly overlapped at the gene level (even when excluding genes not contained in both datasets), further gene set over-representation analysis was performed for pathways specific to each detection approach and the intersection of the two.

The pathway composition of the intersection was further verified by a permutation resampling analysis (as described in Section 2.3.6): the same number of genes detected by SLIPT were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1,652 reactome pathways. These over-representation scores (χ^2) were compared the observed over-representation in the intersection of the SLIPT candidates, with the proportion of resamplings with higher χ^2 values used for empirical p-values of pathway composition. The χ^2 test was used as

an appropriation of Fisher’s exact test on a hypergeometric distribution for resampling to computationally scale pathway over-representation tests across iterations. Pathways for which no resamplings were occurred as high as the observed were reported as $p < 10^{-6}$. These empirical p-values were adjusted for multiple comparisons (FDR). Intersection size was not assumed to be constant across resamplings so similarly with the proportion of resamplings with higher or lower intersection size were used to evaluate significance of enrichment or depletion respectively (of siRNA candidate among SLIPT candidate genes).

2.3 Techniques

Various statistical, computational, and bioinformatics techniques were performed throughout this thesis. This section describes these techniques and gives the parameters used unless otherwise specified. Where relevant, the R package implementation which provided the technique will be acknowledged.

2.3.1 Statistical Procedures and Tests

As described in sections 2.3.4 and 2.2.3, the z-transform has been used to generate z-scores in various analyses in this thesis. Each row of dataset (x) is transformed into a scores (z) using the mean (μ) and standard deviation (σ) of the data such that:

$$z = \frac{x - \mu(x)}{\sigma(x)}$$

This generates data where each row (gene) has a mean of 0 and standard deviation of 1. Where plotted as a heatmap, any data more than 3 standard deviations above or below the mean is plotted as 3 or -3 respectively.

Empirical Bayes differential expression analysis was performed using the `limma` R package (Ritchie *et al.*, 2015). Where specified, the Fisher’s exact test, χ^2 test, and correlation were used to measure associations between variables (as implemented in the `stats` R package (R Core Team, 2016)). Unless otherwise specified, Pearson’s correlation was used for correlation analyses (r) and coefficient of determination (R^2). Where these comparisons are discussed in more detail, Fisher’s exact test and χ^2 tests are supported by a table or Venn diagram, rendered with the `limma` R package (Ritchie *et al.*, 2015). In some analyses, correlation is further supported by a scatter plot and a line of best fit derived by least squares linear regression.

The `t.test` function (R Core Team, 2016) has also been used to implement the t-test to compare pairs of data. Where relevant, an ANOVA has been performed to report

significance of multivariate predictors of outcomes, or least squares linear regression performed for the adjusted coefficient of determination (R^2) and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violinplot visualisation (rendered in R).

Multiple comparisons are adjusted by the Benjamini-Hochberg procedure to control the false discovery rate (FDR) unless otherwise specified (Benjamini and Hochberg, 1995). This procedure adjusts p-values to achieve an average of the proportion of false-positives among significant tests below a threshold, α . The more stringent Holm-Bonferroni (Holm, 1979) was also applied in some cases to adjust for multiple comparisons and control the family-wise error rate which adjusts p-values so that the probability that any one of the tests is a false-positive (type-1 error) below a threshold, α .

2.3.2 Gene Set Over-representation Analysis

Gene set enrichment over-representation was performed to test whether there is an enrichment of a gene set (e.g., a biological pathway) among a group of input genes. Such input genes may be predicted synthetic lethal candidates or a subset defined by clustering (in Section 2.3.3) or comparison with experimental candidates (see Section 2.2.3.1). Initially, these tests were performed using the GeneSetDB web tool (Araki *et al.*, 2012) hosted by the University of Auckland on the Reactome pathways (Croft *et al.*, 2014). Since the GeneSetDB tool used an older version of Reactome (version 40), it was difficult to directly compare with the results of other analysis (see sections 2.2.3.1 and 2.3.6) performed on version 52 (as described in Section 2.1.1.2). Thus an implementation of the hypergeometric test in R (R Core Team, 2016) was used to test for over-representation against Reactome (version 52) pathways. Pathways containing less than 10 genes or more than 500 (as performed in GeneSetDB by Araki *et al.*, 2012) were excluded before adjusting for multiple comparisons.

2.3.3 Clustering

Clustering analysis when performed uses unsupervised hierarchical clustering with complete linkage (distance calculated from the furthest possible pairing). For correlation matrices or multivariate normal parameters (e.g., Σ), the distance metric used was Euclidean distance. For empirical or simulated gene and pathway expression data correlation distance was used, calculated by $distance = 1 - cor(t(x))$ where cor is Pearson's correlation and $t(x)$ is the transpose of the expression matrix.

2.3.4 Heatmap

Standardised z-scores of the data were used to plot heatmaps on an appropriate scale. Raw (log-scale) read counts or voom normalised counts per gene (as specified) were plotted as normalised z-scores on a $[-3, +3]$ blue-red scale. Similarly, correlations were plotted on a $[-1, +1]$ blue-red scale. These heatmaps were performed using the linkage and distance specified for the clustering performed in Section 2.3.3. The `gplots` R package (Warnes *et al.*, 2015) was used to generate many of the heatmaps throughout this thesis, along with a customised heatmap function (released as `heatmap.2x`). Where clearly specified, data have been split into subsets with clustering performed separately on each subset with these plotted alongside each other.

2.3.5 Modeling and Simulations

Statistical modeling and simulations have been used to test various synthetic lethal detection procedures on simulated data. This involves constructing a statistical model of how synthetic lethality would appear in (continuous normally distributed) gene expression data. Where presented (in Section 3.2.1), the assumptions of the model are stated clearly. The model allows sampling from a multivariate normal distribution (using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016)) to generate simulated data with known underlying synthetic lethal partners (detailed in Section 3.2.2). We can test whether statistical procedures, including those developed in this thesis (presented in Section 3.1), are capable of detecting them upon this simulated data. This multivariate normal simulation procedure also enables the inclusion of correlation structure which is either given as correlated blocks of genes or derived from pathway structures (as detailed in Section 3.4.2).

If this multivariate normal distribution was sampled once and the procedure to add known synthetic lethal partners was performed, it generates a simulated dataset. Performing this simulation procedure and testing with a synthetic lethal detection procedure iteratively, these simulations can be used to assess the statistical performance of the detection procedure. The number of iterations (`Reps`) will be given for each simulation result. Typically, these are performed 1000 or 10,000 times depending on computational feasibility of doing so on larger datasets.

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for χ^2 , significance and directional criteria met for SLIPT (see Section 3.1), and an arbitrary threshold: < -0.2 and $> +0.2$ for negative correlation and correla-

tion respectively. Sensitivity (or “true positive rate”) was measured as the proportion of known synthetic lethal partners predicted to be synthetic lethal. Specificity (or “true negative rate”) was measured as the proportion of known non-synthetic lethal partners predicted not to be synthetic lethal. The “false discovery rate” (also used in adjusting for multiple comparisons) was measured here as the proportion of known non-synthetic lethal partners out of all putative partners predicted by the detection procedure. Statistical “accuracy” is the proportion of true predictions for a detection procedure, which is both the correctly predicted known synthetic lethal partners and correctly negative known non-synthetic lethal partners.

2.3.5.1 Receiver Operating Characteristic (Performance)

A more general procedure to measure the statistical performance of a simulation is the Receiver Operating Characteristic (ROC) curve which does not assume a threshold for classification of synthetic lethality but demonstrates the trade-off of sensitivity and specificity (Akobeng, 2007; Fawcett, 2006; Zweig and Campbell, 1993). These curves (implemented with the `ROCR` R package (Sing *et al.*, 2005)) plot the True Positive Rate (sensitivity) against the False Positive Rate (1–specificity) as the prediction threshold is varied. An ideal detection method will have a true positive rate of 1 and a false positive rate of 0, hence the Area Under the ROC curve (AUC or AUROC) is a measure of statistical performance for a detection procedure accounting for this trade-off. AUROC values are typically range from 0.5 the value expected by random chance to 1 for an optimal detection method, however it is possible for an AUROC below 0.5 for a poor detection method that performs worse than random chance. In cancer biology, an AUROC of approximately 0.8 is a predictive biomarker suitable for publication (Hajian-Tilaki, 2013) but predictors with lower AUROC values may still be informative depending on the context. In this thesis, the AUROC values varies widely across simulation parameters and a primarily used for comparisons across these parameters, although they can also be used to refine thresholds for optimal classification.

2.3.6 Resampling Analysis

Resampling analyses (e.g., “permutation” analysis) are used to statistically test the significance of an observation without assuming the underlying distribution of expected test statistics Collingridge (2013). Instead these are derived from randomly shuffling test statistics or randomly sampling predicted candidates. For the purposes of this thesis, this involved randomly sampling genes from those tested to be analysed as putative synthetic lethal candidates. This was performed both for testing the significance

of pathway composition in the intersection with experimental gene candidates (Section 2.2.3.1) and for assessing the significance of pathway structure among synthetic lethal candidates (Section 3.4.1.1).

These were analysed to compare the observed synthetic lethal genes against values derived from randomly sampling the same number of genes as observed by synthetic lethal from among the genes tested. Sampling iteratively across many resampling procedures, these resampling-based values form a null distribution that would be observed if the null hypothesis were true. Thus the proportion of resampling-based values across these iterations that are greater than or equal to that observed, forms an empirically derived p-value to test significance.

Resampling was performed for comparison (in Section 2.2.3.1) with fixed experimental screen candidates (Telford *et al.*, 2015) both resampling the number of genes overlapping with the screen candidates and test statistics for pathway enrichment. Resampling analysis was also applied to shortest paths and network metrics (in Section 3.4.1.1) to test significance of directional relationships between synthetic lethal candidate genes within pathway structures.

The number of iterations determines the accuracy of these p-values. For pathway composition (in Section 2.2.3.1), a million iterations were performed using high performance computing (as detailed in Section 2.5.3) to provide sufficient accuracy after adjusting for multiple comparisons across pathways. For the purposes of network analysis (in Section 3.4.1.1), a thousand iterations were sufficient to reject the null hypothesis for the majority of pathways tested before adjusting for multiple comparisons, and thus further iterations were not performed.

2.4 Pathway Structure Methods

2.4.1 Network and Graph Analysis

Networks are important in considering the structure of relationships in molecular biology, including gene regulation, kinase cellular signalling, and metabolic pathways (Barabási and Oltvai, 2004). Network theory is an interdisciplinary field which combines the approaches of computer science with the metrics and fundamental principles of graph theory, an area of pure mathematics dealing with relationships between sets of discrete elements. The vast amounts of molecular and cellular data from high-throughput technologies have enabled the application of network-based and genome-wide bioinformatics analysis to examine the complexity of a cell at the molecular level

and understand aberrations in cancer. This thesis uses various metrics and analysis procedures developed in Graph and Network theory to analyse graph structure of biological pathways. Where feasible, these have been implemented using the `igraph` R package with such procedures described below (Csardi and Nepusz, 2006). Custom R functions to perform more complex analysis and visualisation of iGraph data objects will be described later.

Graph theory is a branch of pure mathematics which deals with the properties of sets of discrete objects (referred to as a ‘node’ or ‘vertex’) with some pairs are joined (by a ‘link’ or an ‘edge’). While a seemingly reductionist abstraction to mathematically study relationships, graph theory serves has applications in a wide range of studies including life sciences. Network theory is the sub-discipline of graph theory which deals with networks which has become popular due to the vast potential for applications of networks (van Steen, 2010).

Applications vary depending on the situation modelled, particularly in how the edges between vertices are defined, whether they are directed or weighted, and whether multiple redundant edges between a pair of vertices (referred to as ‘parallel edges’) or edges connecting a vertex to itself (referred to as ‘loops’) are permitted in the model. Networks are defined such that the edges represent a relationship between the vertices and may be directed, weighted, or contain parallel edges or loops depending on the application (van Steen, 2010). Unless otherwise stated, graph structures and networks in thesis will be unweighted and have no parallel edges or loops. Where a directional relationship is known or modelled, it will be represented with a directed edge in a directed graph.

2.4.2 Sourcing Graph Structure Data

Pathway Commons interaction data was sourced using Biological pathway exchange (BioPAX) with the paxtools-4.3.0 Java application on October 6th 2015 (Cerami *et al.*, 2011; Demir *et al.*, 2013). This utility was used to source ‘sif’ format interaction data into R (R Core Team, 2016), from which the human Reactome (version 52) dataset of interactions was imported (Croft *et al.*, 2014), matching those used for pathway enrichment analysis. These interactions were used to construct an adjacency matrix for the Reactome network and subnetworks corresponding to each relevant biological pathway.

2.4.3 Constructing Pathway Subgraphs

Subgraphs for each relevant pathway were constructed by matching the nodes in the complete Reactome network to the pathway gene sets (as derived in Section 2.1.1.2). A subgraph with adjacent nodes was constructed by adding nodes which have an edge with a gene in the pathway gene set. The pathways these adjacent nodes belong to were added to form a “meta-pathway” to account for the possibility for nodes within the pathway being linked by the surrounding graph structure.

2.4.4 Network Analysis Metrics

The existing network analysis measures applied in this thesis (as described below) used an implementation in the `igraph` R package where it was available (Csardi and Nepusz, 2006). Otherwise, custom features were developed for analysis of iGraph objects in R and released as `igraph.extensions` (as described in Section 3.5.3).

Vertex degree is the number of edges a node has and is a fundamental measure of the importance and connectivity of a network (van Steen, 2010). More connected nodes, such as network hubs, will have a higher vertex degree relative to other nodes. For the purposes of this thesis, vertex degree ignored edge direction with loops (edges with itself) and double edges to the same node excluded.

A fundamental concept in network analysis is a “shortest path”, that is the shortest route via edges between any two particular nodes in a network. These are computed by Dijkstra’s algorithm (Dijkstra, 1959) in the `igraph` R package (Csardi and Nepusz, 2006). Where applicable paths will only use directed edges in a particular direction. Shortests paths are a useful measure of how close nodes are in a network. This is used to compute information centrality, and for further analysis of pathway structure (as described in Section 3.4.1).

Network centrality is an alternative measure of the importance or influence of a node to the graph structure (Borgatti, 2005). Various strategies are used to derive centrality, typically based on how connected the node is or the impact of node removal on the connectivity of the network. One of the most notable is the “PageRank” algorithm, a refinement of eigenvector centrality based on the eigenvectors of the adjacency matrix (Brin and Page, 1998). This is implemented in the `igraph` R package (Csardi and Nepusz, 2006).

Another network centrality measure that has been previously applied to biological protein interaction networks (Kranthi *et al.*, 2013) is the “information centrality”. The information centrality of a node is the relative impact on efficiency (transmission of

information via shortest paths) of the network when the node is removed. That is the centrality (C) (Kranthi *et al.*, 2013) for node n in graph G is defined as:

$$C_n = \frac{E(G) - E(G')}{E(G)}$$

where G' is the subgraph with the node removed and E is the efficiency (Latora and Marchiori, 2001) derived from shortest paths (d_{ij} between nodes i and j) as:

$$E(G) = \frac{2}{N(N-1)} \sum_{i < j \in G} \frac{1}{d_{ij}}$$

The efficiency of the network can be derived from shortest paths implemented in the `igraph` R package and the iterative network centrality computation of each node has been released as an R package (`info.centrality`) and included in the `igraph.extensions` package.

2.5 Implementation

2.5.1 Computational Resources and Linux Utilities

Several computers were used to process and store data during this thesis (as summarised in Table 2.2), running different versions of Linux operating systems, including a personal laptop computer, laboratory desktop machine, departmental server, and the New Zealand eScience Infrastructure Intel Pan high-performance computing cluster (a supercomputer based at the University of Auckland). Each of these systems support a 64-bit architecture. Current workflows on local machines use Elementary OS (based on the Ubuntu versions given in Table 2.2) and interacting with these via ZSH shell. However, Ubuntu OS and the Bourne Again SHell (bash) were used at the inception of this project and bash is continues to be used for running scripts. Various Linux applications and command-line utilities were used on these machines (as summarised in Table 2.3). As such, the workflows developed in this project should be backwards-compatible with Ubuntu Linux (and other derivatives). The majority of novel methodology and implementations were performed in R which is a cross-platform language, packages developed in R will be available for users of Linux, Mac, and Windows machines.

Table 2.2: Computers used during Thesis

	Viao Laptop	Lab Machine	Biochem Server	NeSI Pan Cluster
Operating System (OS)	Elementary OS Freya 0.3.2	Elementary OS Loki 0.4	Red Hat Enterprise Maipo 7.2	Cent OS Final 6.4
Upstream OS	Ubuntu LTS Trusty 14.04	Ubuntu LTS Xenial 16.04		
Linux Kernel	3.19.0-65-generic	4.4.0-36-generic	3.10.0-327.36.2.el7.x86_64	2.6.32-504.16.2.el6.x86_64
Shell: bash	4.3.11(1)	4.3.46(1)	4.2.46(1)	4.2.1(1)
Shell: zsh	5.0.2	5.1.1	5.0.2	5.2

Table 2.3: Linux Utilities and Applications used during Thesis

	Viao Laptop	Lab Machine	Biochem Server	NeSI Pan Cluster
OS	Elementary OS Freya 0.3.2	Elementary OS Loki 0.4	Red Hat Enterprise Maipo 7.2	Cent OS Final 6.4
Linux Kernel	3.19.0-65-generic	4.4.0-36-generic	3.10.0-327.36.2.el7.x86_64	2.6.32-504.16.2.el6.x86_64
Scripting	Shell bash Shell zsh	4.3.11(1) 5.0.2	4.3.46(1) 5.1.1	4.2.46(1) 5.0.2 5.2
Programming	Python Java C++	2.7.6 1.8.0_101 4.8.4	2.7.12 9-ea 5.4.0	2.7.5 1.8.0.101 4.8.5 4.4.7
Text Editor	nano kile (L <small>A</small> T <small>E</small> X)	2.2.6 2.1.3	2.5.3 2.1.3	2.3.1
Version Control	git	1.9.1	2.11.0	1.7.1 1.8.3.1
Shell Utilities	sed grep nohup	4.4.2 2.16-1 8.21	4.4.2 2.25-1 8.25	4.4.2 2.20 8.22 4.4.1 2.6.3 8.4
Typesetting	TeX TeXLive (L <small>A</small> T <small>E</small> X) PDFTeX pandoc	3.1415926 2013 2.5-1 1.12.2.1	3.14159265 2015 2.6 1.16.0.2	
Remote Computing	Slurm scheduler OpenSSH OpenSSL rsync Globus Online Transfer Cisco AnyConnect VPN			16.05.6 7.2p2 1.0.2g 3.1.0p31 3.1.1p31 3.1.05170
Image Processing	Inkscape GIMP ImageMagick	0.48.4 2.8.10 6.7.7.10-6	0.91 2.8.16	

2.5.2 R Language and Packages

The R programming language has been used for the majority of this thesis. Current R installations across the machines used are given in Table 2.4. Local machines currently run the latest version of the R (at the time of writing) and remote machines run the versions and modules as managed by the system administrator.

Various scripts and packages in this thesis were developed or run in previous versions of RStudio and R but these run without error in the current version of R (and the older

versions on remote machines). The R packages which were used throughout this thesis (as detailed in Table 2.5 with versions specified) were installed from the Comprehensive R Archive Network (CRAN, 2017), Bioconductor (Gentleman *et al.*, 2004, version 3.4; BiocInstaller 1.24.0), or GitHub. These packages were not updated when they would change the functionality of scripts or functions in packages, in particular imported data from annotation packages (used to define gene sets) have been saved as local files to continue using stable versions of these pathway data (across machines).

This is a summary of the key packages which (in addition to their dependencies) have been used throughout this project. Where a package implementation has been central to the methods applied, they are described in more detail in the relevant section. A full table of packages used in this thesis can be found in Appendix B (Table B.1). The R packages developed during this thesis are given in Table 2.6 with the relevant sections describing their implementation and use where appropriate, in addition to further details on these functions in Section 3.5.

Table 2.4: R Installations used during Thesis

	Viao Laptop	Lab Machine	Biochem Server	NeSI Pan Cluster
OS	Elementary OS Freya 0.3.2	Elementary OS Loki 0.4	Red Hat Enterprise Maipo 7.2	Cent OS Final 6.4
Programming	R 3.3.2	3.3.2	3.3.1	3.3.0-intel (module)
Development	RStudio 1.0.136	1.0.136	1.0.136 (server)	

Table 2.5: R Packages used during Thesis

Package	Version Used	Built	Repository
colorspace	1.3-2	3.3.1	CRAN
curl	2.3	3.3.1	CRAN
data.table	1.9.6	3.3.1	CRAN
dendextend	1.4.0	3.3.2	CRAN
DBI	0.5-1	3.3.1	CRAN
devtools	1.12.0	3.3.1	CRAN
dplyr	0.5.0	3.3.1	CRAN
ggplot2	2.2.1	3.3.1	CRAN
git2r	0.18.0	3.3.1	CRAN
gplots	3.0.1	3.3.1	CRAN
gtools	3.5.0	3.3.1	CRAN
igraph	1.0.1	3.3.1	CRAN

matrixcalc	1.0-3	3.3.1	CRAN
mclust	5.2.2	3.3.1	CRAN
mvtnorm	1.0-6	3.3.1	CRAN
org.Hs.eg.db	3.1.2	3.1.2	Bioconductor
openssl	0.9.6	3.3.1	CRAN
plyr	1.8.4	3.3.1	CRAN
purrr	0.2.2	3.3.1	CRAN
reactome.db	1.52.1	3.2.1	Bioconductor
RColorBrewer	1.1-2	3.3.1	CRAN
Rcpp	0.12.9	3.3.1	CRAN
ROCR	1.0-7	3.3.1	CRAN
roxygen2	6.0.1	3.3.2	CRAN
shiny	1.0.0	3.3.1	CRAN
snow	0.4-2	3.3.1	CRAN
testthat	1.0.2	3.3.2	CRAN
tidyverse	1.1.1	3.3.2	GitHub (hadley)
sm	2.2-5.4	3.3.1	CRAN
Unicode	9.0.0-1	3.3.2	CRAN
vioplot	0.2	3.3.1	CRAN
viridis	0.3.4	3.3.2	CRAN
xml2	1.1.1	3.3.2	CRAN
xtable	1.8-2	3.3.1	CRAN
zoo	1.7-14	3.3.1	CRAN
graphics	3.3.2	3.3.2	base
grDevices	3.3.2	3.3.2	base
cluster	2.0.5	3.3.1	base
graphics	3.3.2	3.3.2	base
grDevices	3.3.2	3.3.2	base
Matrix	1.2-8	3.3.1	base
stats	3.3.2	3.3.2	base

Table 2.6: R Packages Developed during Thesis

Package Name	Description and GitHub Repository	Section
<code>slipt</code>	Synthetic lethal detection by SLIPT (to accompany publication) https://github.com/TomKellyGenetics/slipt	3.1
visualisation	<code>vioplotx</code> Customised violin plots (based on <code>vioplot</code>) https://github.com/TomKellyGenetics/vioplotx	2.3.4
	<code>heatmap.2x</code> Customised heatmaps (based on <code>gplots</code>) https://github.com/TomKellyGenetics/heatmap.2x	
igraph.extensions	<code>igraph.extensions</code> Meta-package to install the follow iGraph functions https://github.com/TomKellyGenetics/igraph.extensions	3.5.3
	<code>plot.igraph</code> Custom plotting of directed graphs https://github.com/TomKellyGenetics/plot.igraph	2.4.4
	<code>info.centrality</code> Computing information centrality from network efficiency https://github.com/TomKellyGenetics/info.centrality	3.4.2
	<code>pathway.structure.permutation</code> Testing pathway structure with resampling analysis https://github.com/TomKellyGenetics/pathway.structure.permutation	3.4.1.1
	<code>graphsim</code> Generating simulated expression from graph structures https://github.com/TomKellyGenetics/graphsim	3.4.2

2.5.3 High Performance and Parallel Computing

Another enabling technology for bioinformatics is parallel computing, performing independent operations in separate cores: this “multithreading” is widely used to increase the time to compute results. Bioinformatics is particularly amenable to this since performing multiple iterations of a simulation or testing separate genes is often “embarrassingly parallel”, being completely independent of the results of each other. As such parallel computing is offered by many high-performance “supercomputers” including national research infrastructure.

The New Zealand eScience Infrastructure (NeSI) is a High Performance Computing (HPC) organisation providing the Intel Pan cluster hosted by the University of Auckland (NeSI, 2017). The Pan cluster used throughout this thesis project to optimise and perform computations which would have otherwise been infeasible in the timeframe of thesis. Such technological developments and infrastructure initiatives have enabled bioinformatics research including this project. High performance computing on the Pan cluster was used extensively in this project including for resampling analysis (in sections 2.3.6 and 3.4.1.1), calculating information centrality (in Section 2.4.4), and in simulations (in sections 2.3.5, 3.2, and 3.4.2)

Scripts and data were transferred between the Pan cluster and University of Otago computing resources by `rsync` or the Globus file transfer service (Globus, 2017). R scripts (R Core Team, 2016) were run in parallel with the “simple network of workstations” `snow` R package Tierney *et al.* (2015). This utilised the “message passing interface” (Yu, 2002) when it was feasible with memory requirements to run in par-

allel across multiple compute nodes, otherwise SOCKS was used to access multiple cores within an instance of R and pass input data to them. R jobs were submitted to queue for available resources and run on the Pan cluster via the Simple Linux Utility for Resource Management (Slurm) workload manager (Slurm, 2017). When running R scripts across many parameters or for memory-intensive jobs, Slurm array job submission and independent submission of different parameters via shell commands with arguments passed to R. In some cases, this submission was automated across a range of parameters with Bash scripts.

Chapter 3

Methods Developed During Thesis

In this Chapter, I will outline the rationale and development of various methods used throughout this thesis to examine synthetic lethality in gene expression data, graph structures, models and simulations. First by describing the Synthetic Lethal Interaction Prediction Tool (SLIPT), a bioinformatics approach to triage of synthetic lethal candidate genes. This is considered one of the main research outputs of the thesis, which is supported by comparisons to an experimental screen from a related project and performance on simulated data. These supporting data will be covered in further chapters but preliminary data to support the use and design of SLIPT are provided alongside description of the method. This includes the construction of a statistical model of synthetic lethality in (continuous multivariate Gaussian) gene expression data, which enables testing SLIPT upon simulated data with known synthetic lethal partners. Another key component of the simulation pipeline used later is the generation of simulated data from a known graph structure or simulated biological pathway. The development of this simulation procedure and other statistical treatment of graph and network structures will also be covered. Various R packages have been developed to support this project, most notably the `slipt` package to implement the SLIPT methodology. The additional R packages for handling graph structures, simulations, and custom plotting features will also be described as research outputs of this thesis, methods applied throughout, and contributions to the open-source software community that made this project feasible.

3.1 A Synthetic Lethal Detection Methodology

The SLIPT methodology identifies gene expression patterns consistent with synthetic lethal interactions between a query gene and a panel of candidate interacting partners.

Gene expression is called low, medium, or high by separating samples into tertiles (3-quantiles) for each gene. Genes with insufficient expression across all samples were excluded by requiring that the first tertile of raw counts is above zero. Then a χ^2 test is performed between the query gene and each candidate partner, with the p-values for the χ^2 test being corrected for multiple testing using false discovery rate (FDR) error control to reduce false positives for large candidate gene panels (Benjamini and Hochberg, 1995). Significance was called only if FDR adjusted p-values were below the threshold $p < 0.05$. A synthetic lethal interaction is predicted (as shown in Figure 3.1) when (i) the χ^2 test is significant; (ii) observed low-query, low-candidate samples are less frequent than expected; and (iii) observed low-query, high-candidate and high-query, low-candidate samples are more frequent than expected.

The synthetic lethal prediction procedure has also been adapted to utilise somatic mutation data for the query gene. This is intended to utilise a query gene known to be recurrently mutated in the disease (and dataset), with the majority of mutations inactivating gene function (e.g., null or frameshift mutations). A synthetic lethal in-

		Candidate Gene		
		Low	Medium	High
Query Gene (e.g. <i>CDH1</i>)	Low	Observed less than expected		Observed more than expected
	Medium			
	High	Observed more than expected		

Figure 3.1: Framework for synthetic lethal prediction. SLIPT was designed to identify candidate interacting genes from gene expression data using the χ^2 test against a query gene. Samples are sorted into low, medium, and high expression quantiles for each gene to test for a directional shift. A sample being low in both genes of a synthetic lethal pair is unlikely, since loss of both genes will be deleterious, and is expected to be statistically under-represented in a gene expression dataset. We expect a corresponding (symmetric) increase in frequency of sample with low-high gene pairs. Synthetic lethal candidate (exprSL) partners of a gene are identified by running this procedure on all possible partner genes, selecting those with an FDR-adjusted χ^2 p-value of $p < 0.05$, and meeting the directional criteria. Since synthetic lethal genes are partners of each other commutatively, the symmetric direction criteria are all required such that synthetic lethal genes will be predicted to be partners of each other.

		Candidate Gene		
		Low	Medium	High
Query Gene (e.g. <i>CDH1</i>)	Mutation	Observed less than expected		Observed more than expected
	Wild-type	Observed more than expected		

Figure 3.2: **Synthetic lethal prediction adapted for mutation.** SLIPT was also adapted to identify candidate interacting genes using (somatic) mutation data of the query gene in the χ^2 test. Samples are sorted into low, medium, and high expression quantiles for each candidate gene and tested for a directional shift against mutation status of the query gene. A sample having low expression or mutation for the synthetic lethal pair is expected to be unlikely with a corresponding increase in frequency of sample with mutant-high or wild-type-low gene pairs. Synthetic lethal candidate (mtSL) partners of a gene are identified by running this procedure on all possible partner genes, selecting those with an FDR-adjusted χ^2 p-value of $p < 0.05$, and meeting the directional criteria.

teraction is predicted (as shown in Figure 3.2) when (i) the χ^2 test is significant; (ii) observed mutant-query, low-candidate samples are less frequent than expected; and (iii) observed mutant-query, high-candidate and wild-type-query, low-candidate samples are more frequent than expected. Unless otherwise specified, computationally predicted synthetic lethal gene candidates from SLIPT used expression data (exprSL) for both genes (as shown in Figure 3.1) rather than mutation data (mtSL) for the query gene (as shown in Figure 3.2).

The SLIPT methodology is amenable for use on expression data including pathway metagenes (as generated in Section 2.2.3). The suitability of the SLIPT methodology to application on public gene expression data will further be supported by simulation results in Section 3.3 and Chapter 6, including comparison to other statistical methods. SLIPT results for *CDH1* will also compared experimental screen results in a breast cell line (Telford *et al.*, 2015), primary screen results are discussed in Section 4.2 and secondary (validation) screen results are presented in Appendix ??.

3.2 Synthetic Lethal Simulation and Modelling

A statistical model of Synthetic Lethality was developed to generate simulated data to test the SLIPT procedure. This section will describe the synthetic lethal model and the simulation procedure for generating gene expression data with known synthetic lethal partners. Some preliminary results to support usage of the SLIPT methodology throughout this thesis will be presented here. The simulation procedure will be applied in more depth in Chapter 6, including in combination with simulations from graph structures.

3.2.1 A Model of Synthetic Lethality in Expression Data

A conceptual model of synthetic lethality was constructed (see Figure 3.3), which will be used to build a statistical model of synthetic lethal gene expression from which to simulate expression data to on which test SLIPT and various potential synthetic lethal prediction methods. In the model, synthetic lethality arises between genes with related functions as a cell death phenotype when these functions are removed.

This model suggests that synthetic lethality is detectable in measures of gene inactivation across a sample population, namely mutation, DNA copy number, DNA methylation, and suppression of expression. While any of these mechanisms of gene inactivation could lead to synthetic lethality, expression data is readily available and changes in these alternative mechanisms are likely to impact on the amount of expressed (functional) RNA or protein detectable. There are several ways that functional relationships between genes could manifest in expression data, including coexpression, mutual exclusivity and directional shifts. Co-expression is overly simplistic and has previously performed poorly as a predictor of synthetic lethality (Jerby-Arnon *et al.*, 2014), although this will still be tested with correlation measures in later simulations. Here the alternative hypothesis is that synthetic lethality will lead to a detectable directional shift in the number of samples exhibiting low or high expression of either gene. This model does not preclude mutual exclusivity (Wappett *et al.*, 2016), compensating expression or co-loss under-representation (Lu *et al.*, 2015) as previously postulated to occur between synthetic lethal genes.

The first condition of the synthetic lethal model is that if there are only two synthetic lethal genes (e.g., *CDH1* and one SL partner), then they will not both be non-functional in the same sample (in an ideal model). Gene function is thus determined for each sample in a model of synthetic lethal with the proportion of samples with a functional or non-functional gene being arbitrary. Whether a gene is functional can

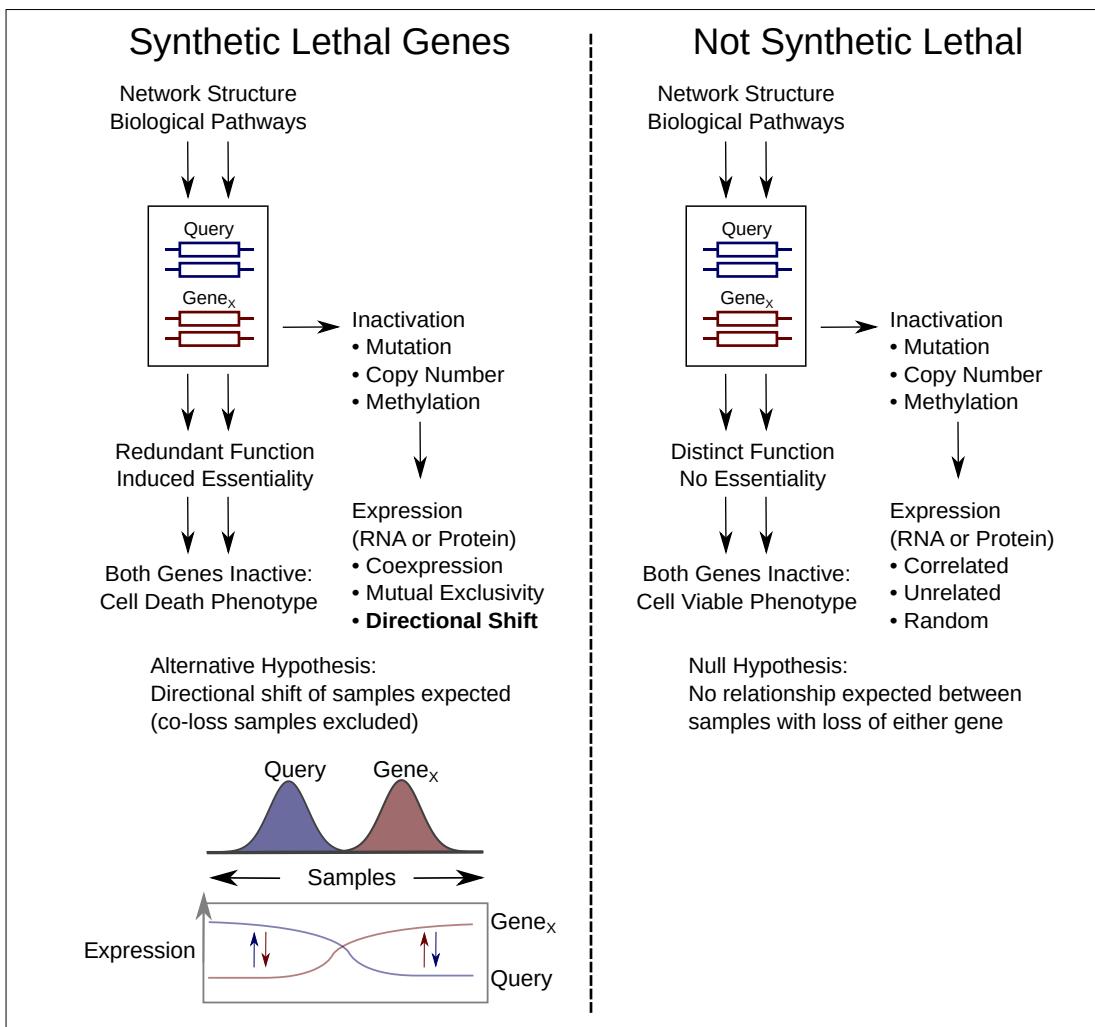


Figure 3.3: A model of synthetic lethal gene expression. A conceptual model of synthetic lethal interactions between a Query gene and partner gene (G_X). Genes that are synthetic lethal may not both be non-functional in the same sample without another gene compensating for the loss of function. This is most likely to be detectable as low gene expression, whether they are lost by mutation, deletion, DNA methylation, or suppressing regulatory signals. This could manifest as coexpression, mutual exclusivity, or directional shifts in sample frequency. Thus the alternative hypothesis (H_A) is that synthetic lethal genes will have a reduced frequency of co-loss samples while the null hypothesis (H_0) is that non-synthetic lethal gene pairs would show no such relationship, even if they may be correlated for other means such as pathway relationships. In this model synthetic lethal genes may compensate for the loss of each other but this is not assumed, only that loss of both is unfavourable to cell viability and probability of detecting samples with combined gene loss.

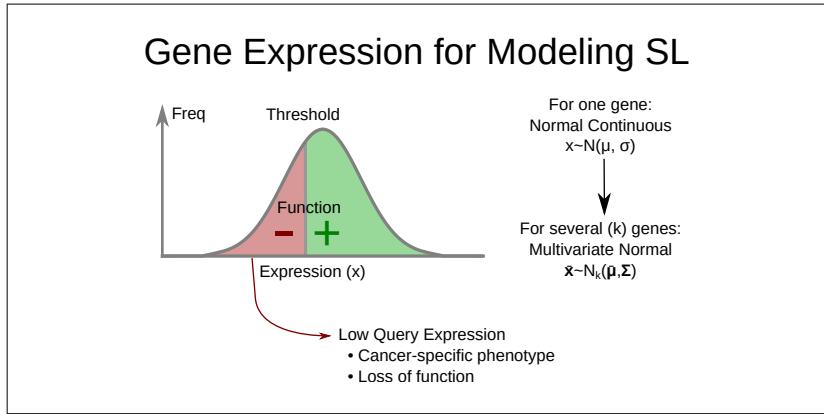


Figure 3.4: Modeling synthetic lethal gene expression. When modeling synthetic lethal interactions between a Query gene with partner genes (G_X and G_Y) above, cellular viability requires that at least one of genes is not inactivated. Expression below a threshold is used as a model of loss of function, where genes are regarded as non-functional for the purposes of modelling synthetic lethality. Tumour suppressor genes with loss of function also have cancer specific phenotypes (although these thresholds are not necessarily the same). Expression is modeled by a normally (Gaussian) distributed continuous data such as (log-scale) data from RNA (microarray or RNA-Seq), protein, or pathway metagenes. This rationale generalises for several genes on a multivariate normal distribution.

similarly be modelled by an arbitrary threshold of continuous and normally distributed gene expression data to define gene function (as shown in Figure 3.4). For the purposes of modeling synthetic lethality in breast cancer expression data, a threshold of the 30th percentile of the expression levels was used because approximately 30% of samples analysed had *CDH1* inactivation. This was generalised for a model of the proportion of samples inactivated for each gene. In this ideal case, no samples lowly expressing both of these genes are expected to be observed. While this is not observed, that is to be expected as it is unlikely that only 2 genes will have an exclusive synthetic lethal partnership. The threshold of the 0.3 quantile was used in simulations derived from this model throughout this thesis.

A synthetic lethal pair of genes is unlikely to act in isolation, therefore higher-order synthetic lethal interactions (i.e., 3 or more genes) must be considered in the model as shown in Figure 3.5. Even when testing pairwise interactions, modelling higher level interactions that may interfere is important. If there are additional synthetic lethal partners, there are two possibilities for adding these: 1) that they are independent partners of the query genes interacting pairwise (and not with each other) or 2) that

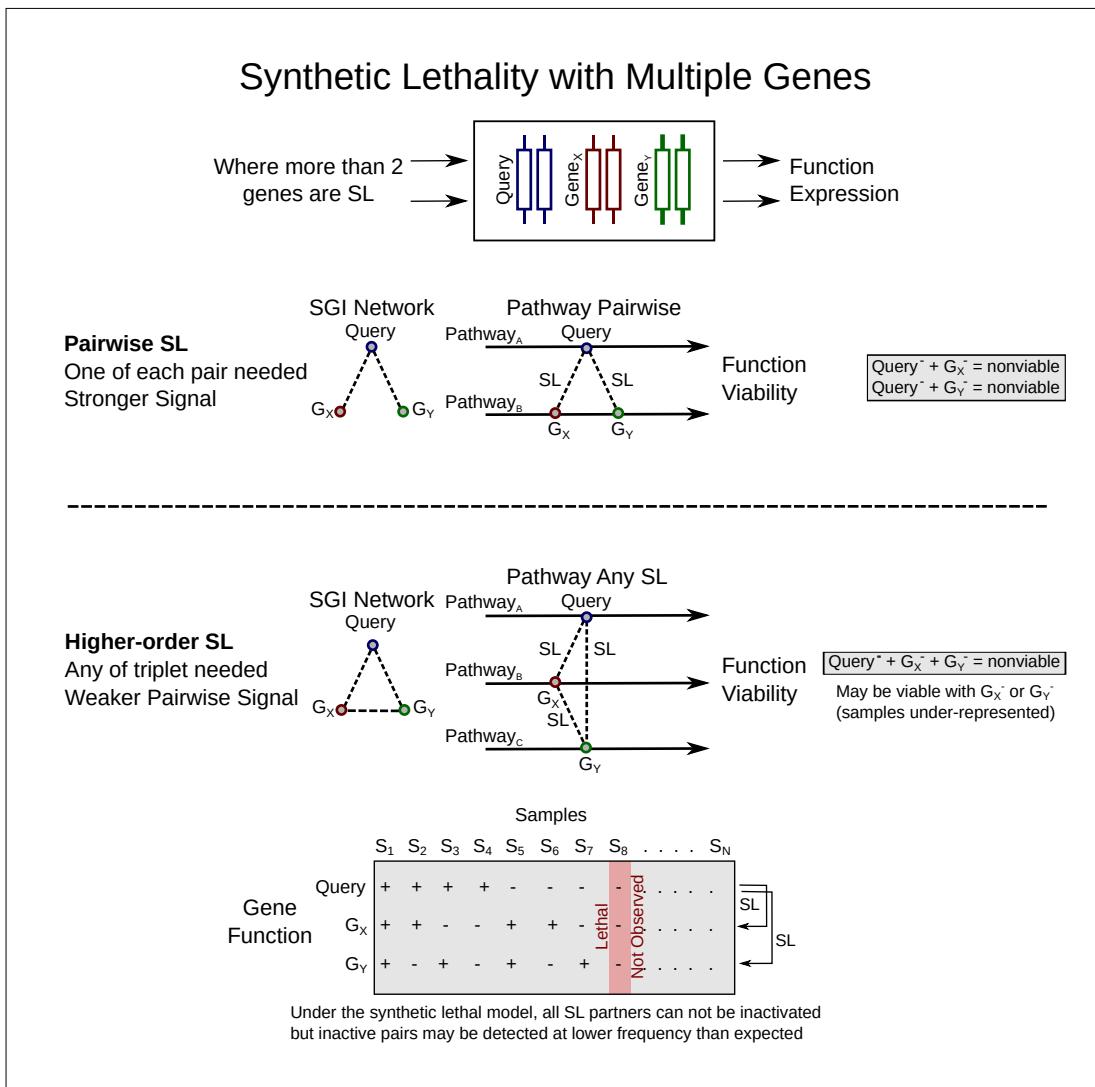


Figure 3.5: **Synthetic lethality with multiple genes.** Higher order synthetic lethal interactions may occur between 3 or more genes, affecting the simulated expression (or synthetic lethal predictions) even if undetected when observed pairwise. Consider interactions between a Query gene and two partner genes (G_X and G_Y). They may interact with the Query pairwise (inviolate when either gene pair is lost) or form a higher-order interaction such as the “synthetic lethal triplet” if any of the genes provide an essential function (inviolate only when all are lost). Either is plausible with the potential pathway structures. A synthetic lethal triple has 8 potential combinations of gene functional but one is not expected to be observed (due to inviability) but pairwise inactivation may be observed if additional partner genes are functional. The proportion of these combinations vary depending on the functional threshold.

an addition partner gene interacts with both of the synthetic lethal genes already in the system and any of the three (or more) are required to be functional for the cell to survive.

The signal (in terms of gene expression data) will be weaker for this latter case and this model has the more stringent assumption that all synthetic lethal partner genes interact with each other: that only one of these must be expressed to satisfy the model of synthetic lethality. In this model any of the synthetic lethal genes in a higher-order interaction is able to provide the missing function of the others, allowing for higher-level synthetic lethal partners to compensate for loss a synthetic lethal gene pair. While samples expressing low levels of the synthetic lethal gene pairs will be under-represented, they may not be completely absent from the dataset due to these higher-level interactions.

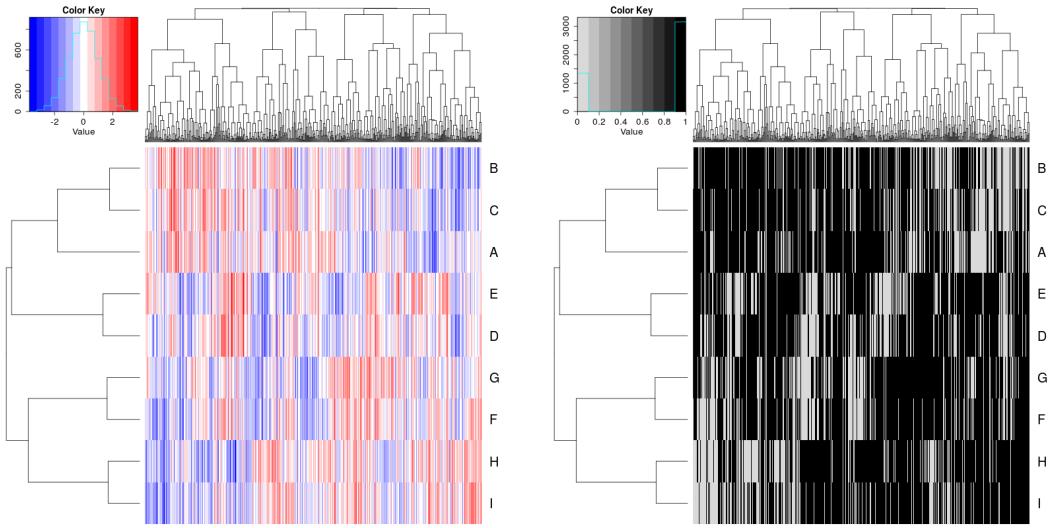
In the example of 3 synthetic lethal genes 3.5, only one of genes involved in the higher-order synthetic lethal interaction is required for cell viability. For synthetic lethal pairs, only a subset of these samples will be inviable (i.e., removed from simulated data), leading to an under-representation.

In practice, samples are not removed from a simulated dataset, rather the expression and function of the query gene is generated across samples separately from the pool of potential partner genes. The query gene data is matched to simulated samples (as shown in Figure 3.7), satisfying the synthetic lethal condition with the procedure described in Section 3.2.2. This is performed to maintain a comparable samples size across simulations and the preserve the assumed (multivariate) normal distribution of the data.

3.2.2 Simulation Procedure

Simulations were developed to simulate normal distributions of expression data and define function with a threshold cut-off. This is the reverse to the procedure of SLIPT to predict synthetic lethal partners (although the threshold is assumed to be unknown when testing upon simulated data). While gene function is used as an intermediary step in modelling synthetic lethal genes in expression data, the normal distribution is sampled for simulated data to represent normalised empirical gene expression data for which SLIPT (and other methods) will be applicable.

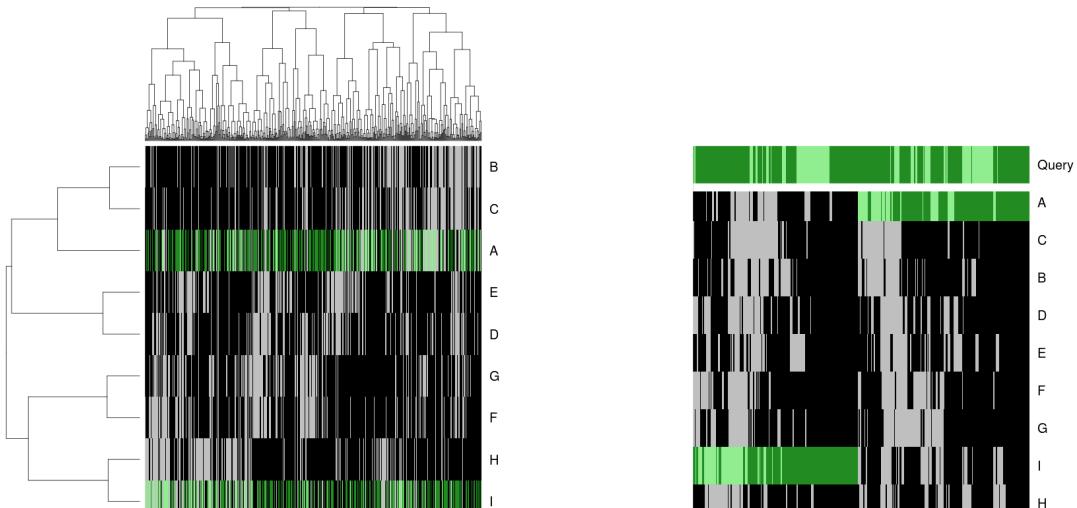
Sampling a distribution for expression profiles has the added advantage of being amenable to simulating correlation structures with the multivariate normal distribution (using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016)). The



(a) Simulated expression matrix

(b) Corresponding gene function calls

Figure 3.6: **Simulating gene function.** A simulated dataset with samples (columns) and genes A–H (rows) is transformed from a continuous (coloured blue–red) scale to a discrete matrix of gene function (black for functional levels and grey for non-functional).



(a) Simulated gene function with SL genes (b) Query gene added with SL condition

Figure 3.7: **Simulating synthetic lethal gene function.** In a discrete simulated gene function dataset (shaded for functional levels and pale otherwise) with samples (columns) and genes (rows), genes A and I are SL partners of a “Query” gene. A partner is selected (highlighted in green) randomly in each sample for simulating synthetic lethality, then ordered such that the query gene or an SL partner are functional in each sample.

parameter Σ is a covariance matrix defines the correlation structure between simulated genes being sampled. With a diagonal of one, this Σ matrix simulates genes with a standard deviation of one and the covariance parameters between them are the correlations between each gene. In Figure 3.6, an example of such a simulated multivariate normal dataset is shown with the functional threshold applied.

Once we have generated a simulated dataset, the samples are compared by gene function (as derived from a functional threshold). Known underlying synthetic lethal partners are selected within the dataset and a query gene is generated by sampling from the normal distribution. These are matched (as shown for 2 synthetic lethal partners in Figure 3.7) such that the synthetic lethal condition is met: that at least one of the synthetic partner genes and the query gene are functional in any particular cell. The samples are ordered by functional data (without assuming correlation of underlying expression values) with the query gene in one direction and the remaining dataset ordered by the selected synthetic lethal partner.

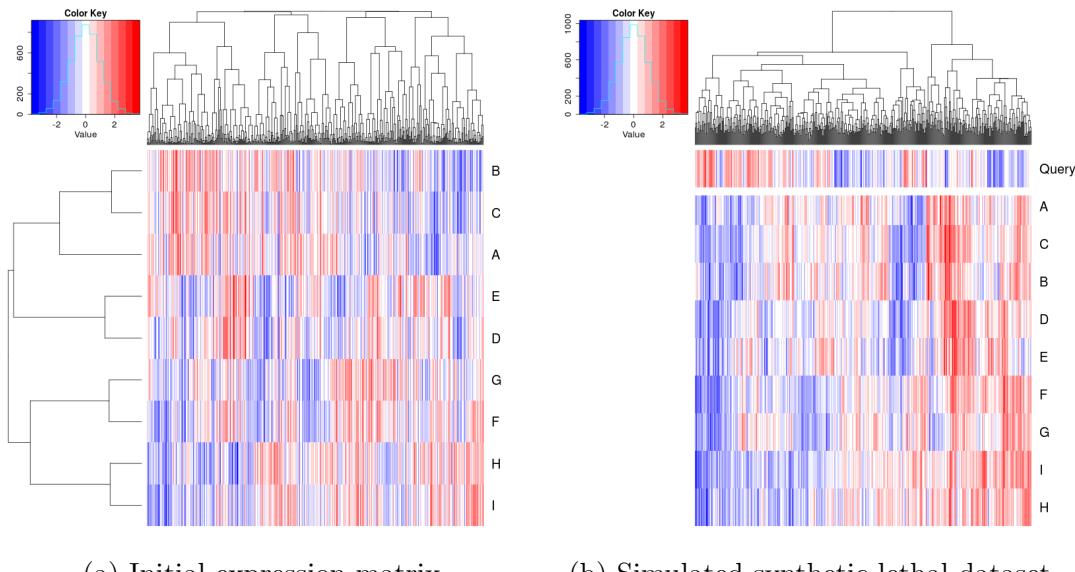


Figure 3.8: Simulating synthetic lethal gene expression. A simulated continuous expression dataset (blue–red scale) with samples (columns) and genes (rows) is matched to a query gene such that at least one synthetic lethal partner is above a functional threshold when the query gene is below it satisfying the synthetic lethal model.

This results a simulated dataset where samples with non-functional query gene have at least one functional partner gene. Similarly, the query gene is functional in all samples where all of the synthetic lethal partner genes are non-functional. Therefore a

dataset has been generated with known synthetic lethal partners (see Figure 3.8) by as few assumptions about the relationships between the each synthetic lethal pair as possible (and allowing compensating functions from higher-order interactions). This has been designed to have the most stringent (least detectable) synthetic lethal relationships where higher-order interactions are possible for the purposes of testing pairwise detection procedures such as SLIPT.

3.3 Detecting Simulated Synthetic Lethal Partners

The synthetic lethal detection methodology (SLIPT), as described in Section 3.1, was tested on simulated data with known synthetic lethal partners, generated using the procedure described in Section 3.2.2. This section will present basic simulations to demonstrate the methodology and support it's use throughout this thesis. These will be performed with sampling from basic statistical distributions as described, including multivariate normal distribution with correlated blocks of genes, with the Σ matrix show in the plots where relevant. A more complex multivariate normal sampling procedure based on pathway graph structures, as described in section 3.4.2, will be applied in Chapter 6.

3.3.1 Binomial Simulation of Synthetic lethality

A previous version of the synthetic lethal simulation procedure (described in Section 3.2.2), used gene function sampled directly from a binomial distribution using the binomial probability of observing functional gene levels ($p = 0.3$) in one observation ($n = 1$) for each samples:

$$X \sim \text{Bin}(n, p)$$

Once a query gene consistent with synthetic lethality has been added, these functional levels were passed directly into SLIPT as “low” and “high” categories.

The simulation procedure was performed with 20,000 total genes (as feasible in the human genome and expression datasets) with a variable number of true synthetic lethal partners and sample sizes of 500, 1000, 2000, and 5000. Each ROC curve was derived from the results of 10,000 replicate simulations. The statistical performance (as shown in Figure 3.9) of such an approach based on the χ^2 p-value declines towards random predictions (an AUROC of 0.5) with an increasing number of underlying true synthetic lethal partners to detect. However, increased sample size mitigates this decline to some extent, as expected with a statistical predictor, particularly for moderate numbers of synthetic lethal partners.

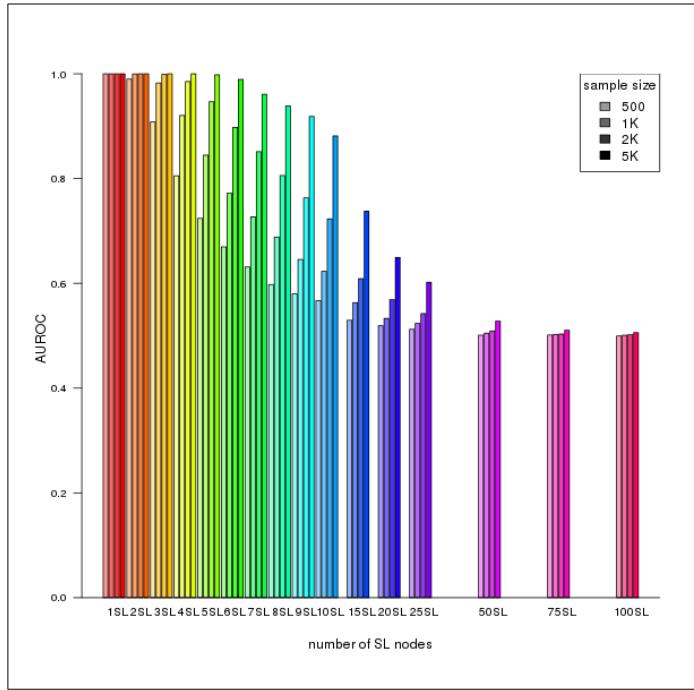


Figure 3.9: Performance of binomial simulations. Gene function was simulated by binomial sampling and tested for synthetic lethal genes. Statistical performance declines with additional known synthetic partners but this is mitigated by increased sample sizes.

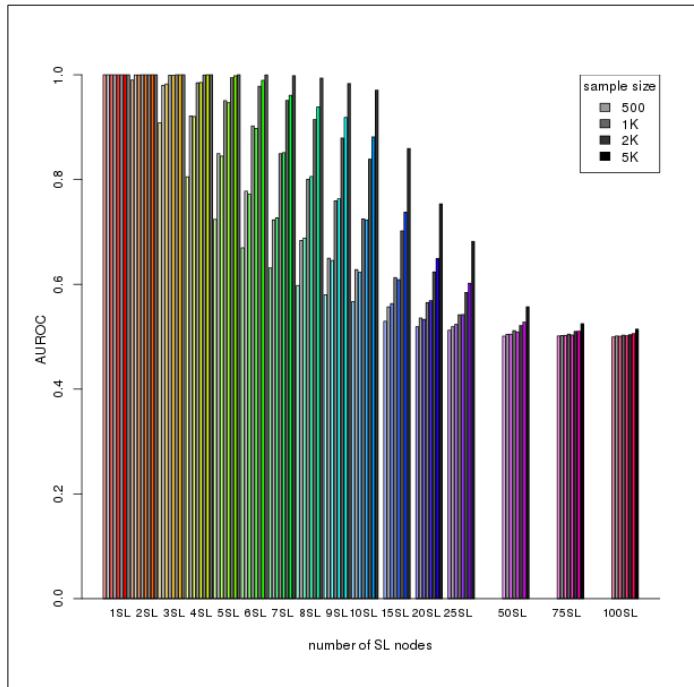


Figure 3.10: Comparison of statistical performance. Binomial simulation of synthetic lethality (in colour) is compared (in greyscale) to multivariate normal simulations (detailed below) which consistently outperforms binomial simulation across parameters.

Simulations based on a simple binomial model of synthetic lethality are limited but form a basis for building a more complex model including expression and correlation structures. While this does not represent the data that SLIPT will be applied to, binomial simulations do demonstrate that SLIPT is able to distinguish small numbers of synthetic lethal partners in a simplistic simulated system with behaviour expected with respect to sample size. This supported further development of the synthetic lethal model and simulation pipeline (as described in Section 3.2) using the multivariate normal distribution.

The multivariate normal simulation procedure is more representative of the (normalised) expression data SLIPT is intended for and enables the prediction procedure to be tested without changes to the methodology (presented in more detail in Section 3.3.2). Sampling continuous expression values from a normal distribution allows the expression threshold for gene function to differ from the categorical “low” and “high” expression binning performed by SLIPT (as discussed in Section 3.2.1) which represents that the SLIPT procedure does not assume a known threshold for expression but rather uses expression as an estimate of gene function. This functionality can be included in the multivariate normal simulation without compromising the statistical performance of the SLIPT, rather the performance estimates (shown in Figure 3.10) were a marked improvement over the binomial simulation procedure across simulation parameters in an equivalent simulation (without correlation structure). This improvement may be due to binomial model defining the synthetic lethal condition in a way that, while ensuring at least one synthetic lethal partner is active in query deficient samples, disrupts the number of samples with functional synthetic lethal genes compared to other genes affecting the expected sample proportions of χ^2 test.

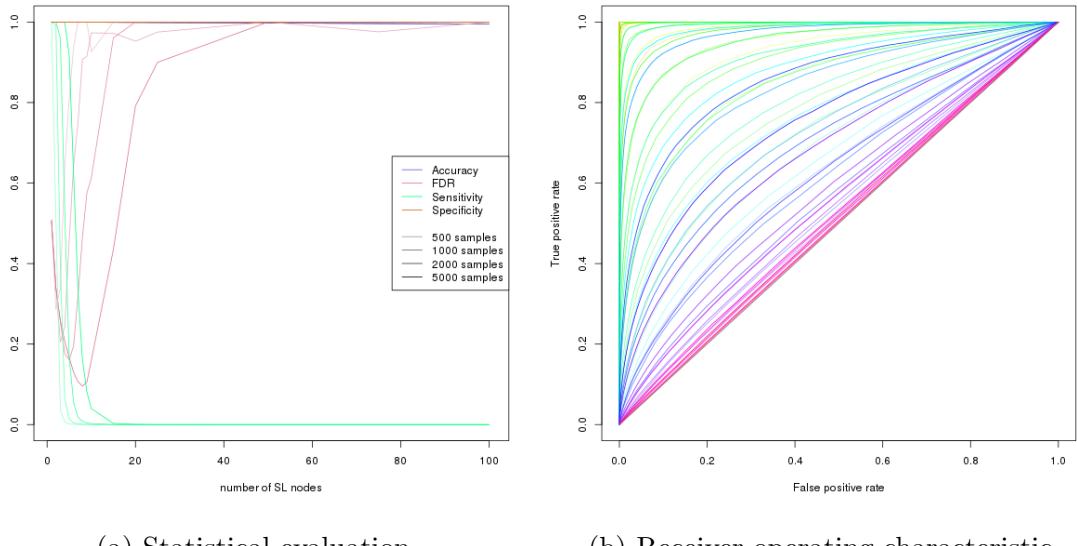
3.3.2 Multivariate Normal Simulation of Synthetic lethality

The multivariate normal simulation procedure was initially performed using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016) (as described in Section 3.2) without correlation structure.

Expression is sampled from multivariate normal distribution with a mean ($\mu = 0$), standard deviation ($\sigma = 1$), and no correlation between genes ($r = 0$):

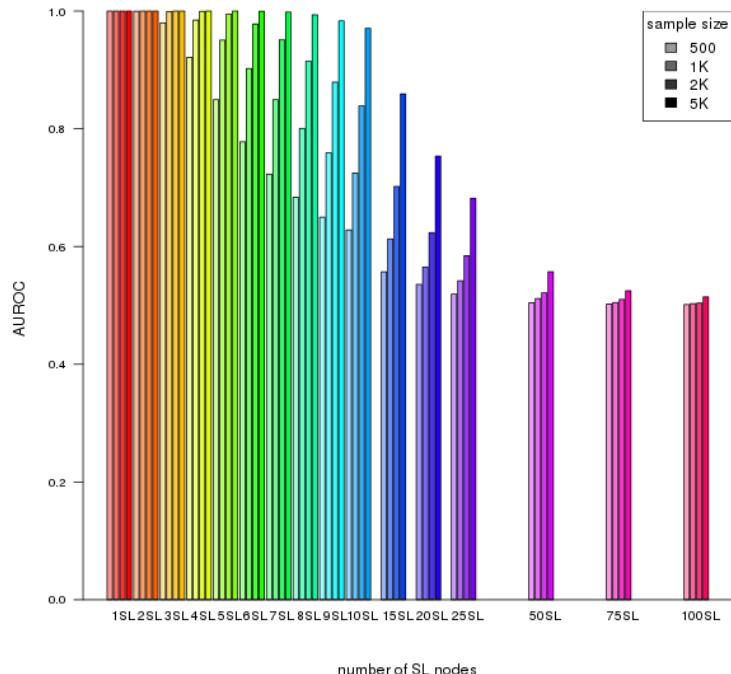
$$X \sim N(\bar{\mu}, \Sigma)$$

Once a query gene consistent with synthetic lethality has been added, the simulated expression values are tested by SLIPT exactly as described in Section 3.1.



(a) Statistical evaluation

(b) Receiver operating characteristic



(c) Statistical performance

Figure 3.11: **Performance of multivariate normal simulations.** Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (without correlation structure). Performance of SLIPT declines for more synthetic partners but this is mitigated by increased sample sizes (in darker colours). This generally occurs as the sensitivity decreases for a greater number of true positives to detect, leading to a trade off in accuracy as seen in a trough for false discovery rate and the ROC curves.

As shown in Figure 3.11a, the statistical accuracy of SLIPT as a binary classifier is considerably high across simulations of a full human dataset of 20,000 genes. However, with the χ^2 p-value as a threshold for prediction, this is largely to desirable specificity: the majority of non-SL genes are distinguished from the few underlying synthetic lethal genes. In this regard, the SLIPT methodology generally performs better with larger datasets with more expected negatives and thus the results of simulations of smaller numbers of genes (e.g., the graph structures analysed in Chapter 6) can be applied to larger datasets where they are expected to perform comparably or better with a lower false negative rate. Accordingly, key results will be supported by replication with larger numbers of non-SL genes added to the simulations.

However, with higher numbers of synthetic lethal genes to detect, the sensitivity (in Figure 3.11a) of SLIPT as a binary classifier of synthetic lethality declines, although this is somewhat mitigated by higher sample sizes (shown in darker colours). Thus the minority of true synthetic lethal partners are more difficult to distinguish when there are more of them (and a weaker expression signal from each). While a reasonable reduction of the false discovery rate can be achieved for moderate numbers of underlying synthetic lethal partners, we can not be sure how many partners are expected to be detected in analyses of expression data. However this simulation procedure is amenable to assessing the performance of SLIPT across simulation parameters, graph structures and comparisons to other approaches (presented in more detail in Chapter 6).

Not all of the genes detected by SLIPT will be true synthetic lethals but these will be among the strongest candidates and it performs better with fewer underlying synthetic lethals to detect. This supports a focus on pathway analyses, in particular detecting pathways for further investigation. Since individually gene candidates are not necessarily gene synthetic lethal themselves, pathway over-representation analysis will be performed to detect functional groups recurrently detected by SLIPT as these detection of functionally related genes further support their role in synthetic lethal relationships in addition to being biologically informative. Alternatively, pathway meta-genes will reduce the number of underlying synthetic lethals to identify synthetic lethal pathways. Both of these approaches will be applied in Chapter 4 to identify and replicate synthetic pathways of *CDH1*. Pathways are also more likely to replicate across experimental models as demonstrated by Dixon *et al.* (2008).

The receiver operating characteristic curves (in Figure 3.11b) demonstrate that SLIPT is subject to near equal trade-off between sensitivity and specificity across threshold values. The lower sensitivity and higher specificity with a binary classification

(in Figure 3.11a) stems from stringent testing by SLIPT with (FDR) p-values adjusted for multiple tests. The area under these curves is also used to compare statistical performance (in Figure 3.11c), with declining performance across increased underlying synthetic lethal partners and increased performance with sample size in multivariate normal simulations.

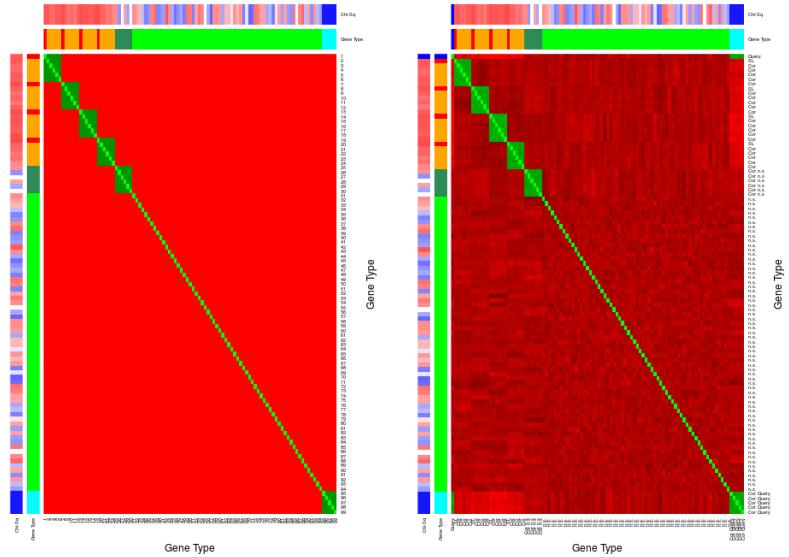
3.3.2.1 Multivariate Normal Simulation with Correlated Genes

Correlation structures can be added to the simulation procedure (as discussed in Section 3.2), starting with simple correlated blocks of genes as the Σ parameter depicted in Figure 3.12a. These correlated blocks represent genes with correlated expression such as that expected by coregulation or biological pathways. Figure 3.12 gives an example of 4 synthetic lethal genes (out of 100), each with 5 correlated genes that are not themselves synthetic lethal partners of the query gene. This serves to test whether synthetic lethal genes are distinguishable from correlated partners. This Σ matrix produces a similar correlation structure (Figure 3.12b) in the resulting expression profiles (Figure 3.12c) where apart from correlated blocks of genes ($r = 0.8$), the remaining genes have only slight variations due to random sampling. The structure of the dataset, particularly between synthetic lethal genes and the query, is shown at the gene expression (Figure 3.12c) and function (Figure 3.12d). These are ordered by the SLIPT results and the synthetic lethal genes are ranked high, with the majority of them being distinguishable from highly correlated genes.

The use of correlation structures generalises to larger datasets, such as 1000 genes shown in Figure 3.13. Synthetic lethal genes are highly ranked by SLIPT and still largely distinguishable from correlated genes. As previously discussed in Section 3.3.2, these synthetic lethal genes are still detectable among a larger number of true negatives and the SLIPT methodology performs better on such datasets.

These plots (Figures 3.12 and 3.13) also show similar correlated blocks with a non-synthetic lethal gene (true negative) and the query gene (which is not synthetic lethal with itself). Neither of these should be synthetic lethal (or detected to be) but they may impact upon the performance of the model, particularly the specificity as correlated negative genes may be distinguishable from true synthetic lethals. The non-synthetic lethal correlated block has no impact on synthetic lethal detection but the impact of query correlated genes will be discussed in Section 3.3.2.2 and Chapter 6.

These simulations (on 100 genes) were repeated to examine the variation between detection on different samples and varying the number of underlying synthetic lethal partners, in simulated gene expression data with correlations structure. A small number



(a) Input Σ matrix parameter (b) Simulated correlation matrix

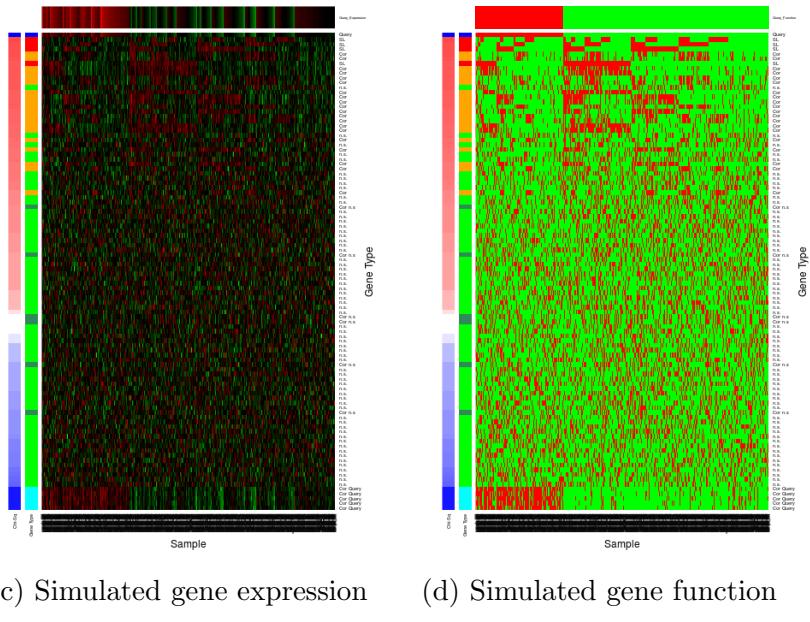


Figure 3.12: Simulating expression with correlated gene blocks. A Σ matrix (a) is used to generate a multivariate normal distribution with 100 genes in correlated blocks of genes (correlated by 0.8) with a comparable structure (b) to the input Σ , as shown by correlation on a red–green scale. The annotation bars for genes give the χ^2 (in blue if the direction of SLIPT is met or red otherwise) and the gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL). The simulated gene expression (c) and function (d) generated are ordered by χ^2 showing the functional structure of synthetic lethal genes and that they are among the strongest SLIPT results.

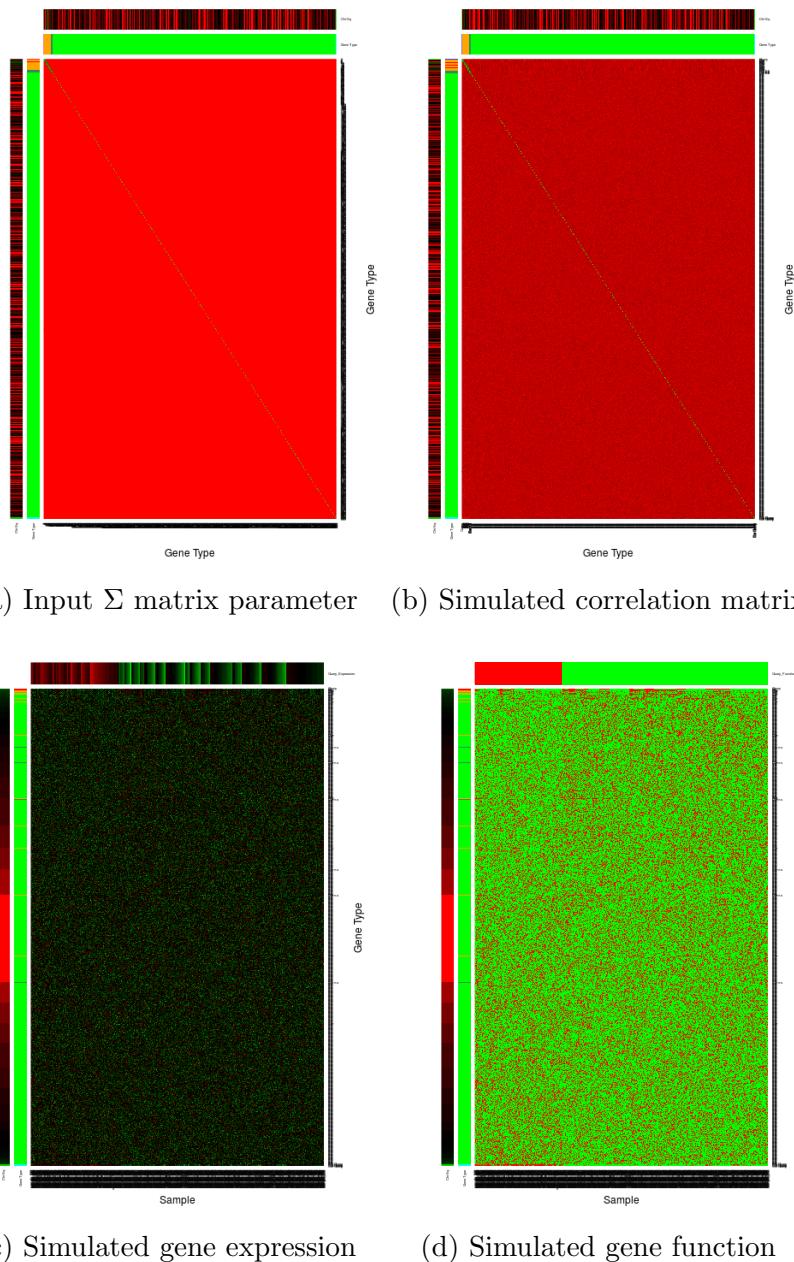
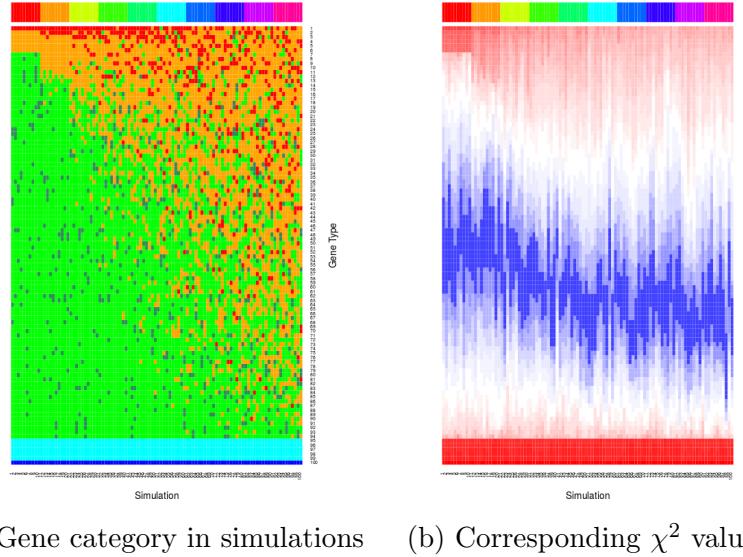


Figure 3.13: Simulating expression with correlated gene blocks. Using the (a) Σ matrix, sampling from a multivariate normal distribution with of 1000 genes produced (b) correlated blocks of genes (correlated by 0.8) on a red-green scale. The simulated gene expression (c) and function (d) generated are ordered by χ^2 and SLIPT direction show that synthetic lethal genes are among the strongest SLIPT results with high specificity against many potential false positives. These are annotated for χ^2 (on a red-green scale) and category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) for each gene.

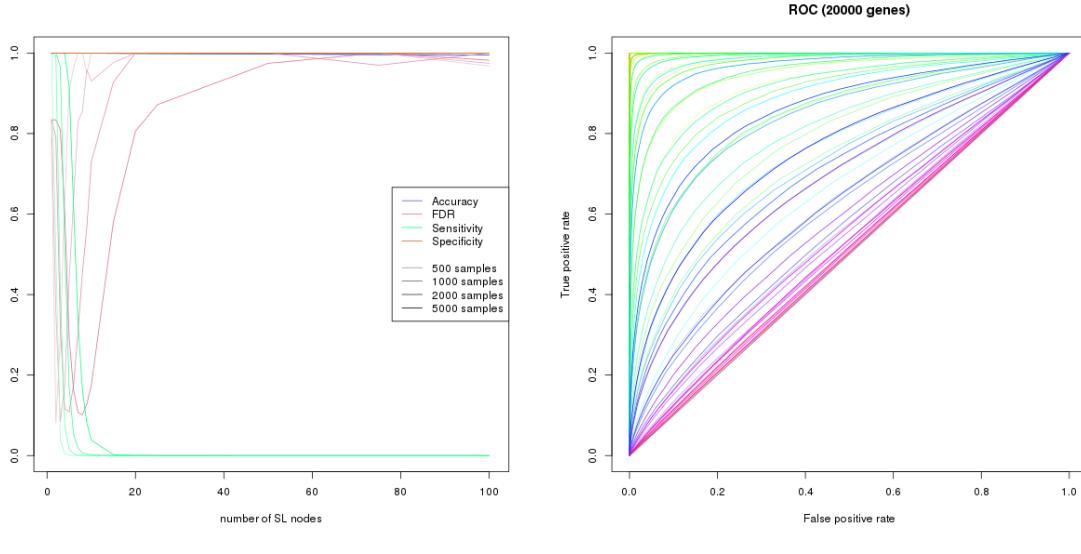


(a) Gene category in simulations (b) Corresponding χ^2 values

Figure 3.14: Synthetic lethal prediction across simulations. The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations. For each of 1–10 SL partners, 10 simulations demonstrate that the increasing numbers of SL partners become harder detect. The χ^2 values show a clear threshold for SL and correlated genes when there are fewer of them, distinguishable from correlated genes in this case.

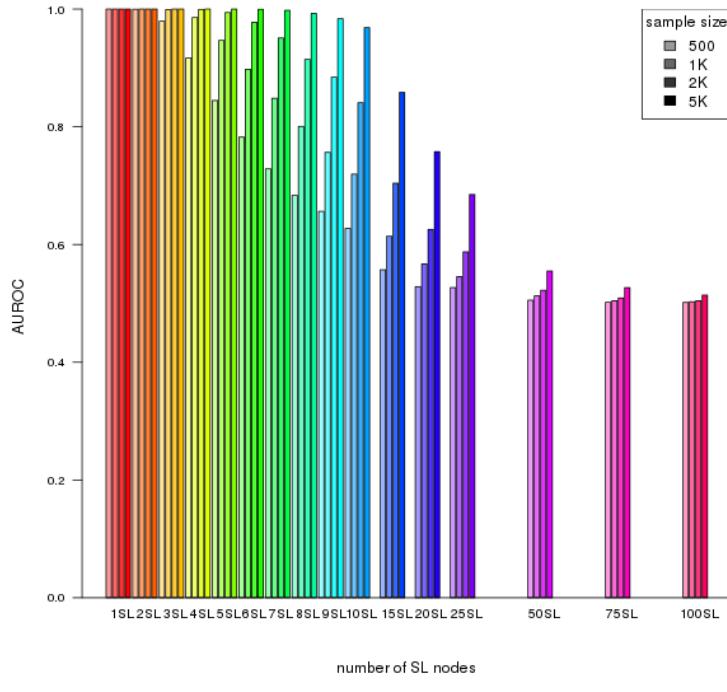
(10 for each) simulations are shown in Figure 3.14 to demonstrate the variation between replicate simulations, with iterative sampling from the same multivariate normal distribution. These simulations show synthetic lethal genes are not only highly ranked by SLIPT when there are few of them but also that they are fairly consistent across replicate simulations. Whereas they become less consistent for increasing numbers of true synthetic lethal partners to detect and thus more difficult to distinguish from other genes, particularly those correlated with them. Similarly, the χ^2 values show a marked stepwise increase with clear thresholds for SL and correlated genes in simple simulations, whereas these become less evident for higher numbers of SL partners.

Whether the synthetic lethal genes detected in simple simulations (in Figure 3.14) are robustly detectable across greater number of simulations, in addition to further comparisons, was tested with a supporting ROC analysis. These results (in Figure 3.15) are very similar to simulations without correlation structure, with SLIPT as a binary classifier having a poor sensitivity with increasing numbers of synthetic lethal partners to detect but high specificity in a total of 20,000 genes with the vast majority being true



(a) Statistical evaluation

(b) Receiver operating characteristic



(c) Statistical performance

Figure 3.15: Performance with correlations. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (with correlation structure). Performance of SLIPT declines for more synthetic partners but this is mitigated by increased sample sizes (darker colours). This generally occurs as the sensitivity decreases for a greater number of true positives to detect, leading to a trade off in accuracy as seen in a trough for false discovery rate and the ROC curves.

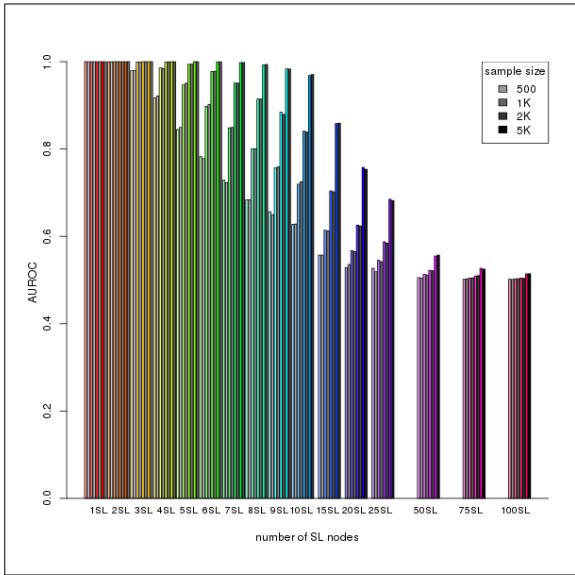
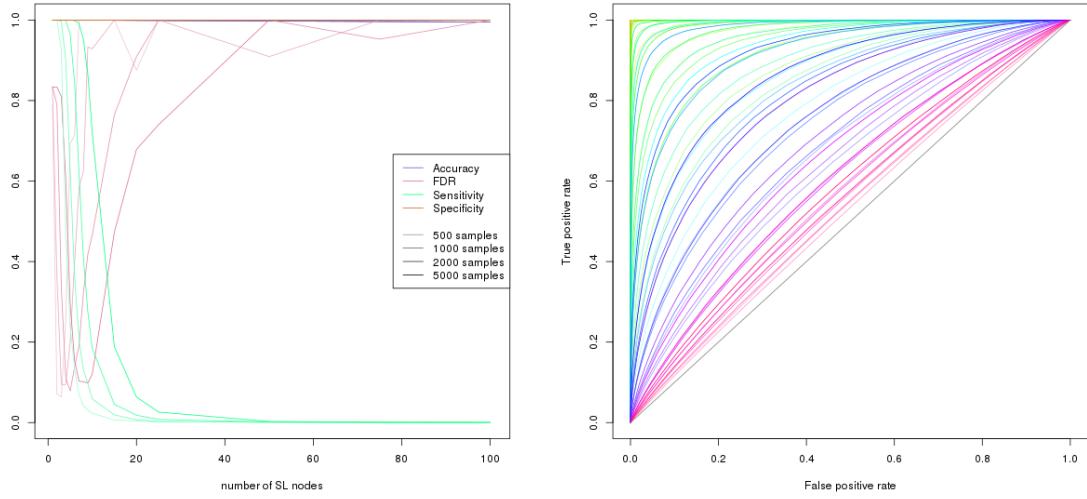


Figure 3.16: Comparison of statistical performance with correlation structure. Multivariate simulation of synthetic lethality with correlation structure (in colour) has comparable performance to simulation without correlations (in greyscale) with known synthetic partners across parameters.

negatives. This is reflected in a similar decline in statistical performance for increasing numbers of synthetic lethal partners and a compensating increase in performance with higher sample size. Overall, the statistical performance is very similar to simulations without correlation structure (as shown in Figure 3.16).

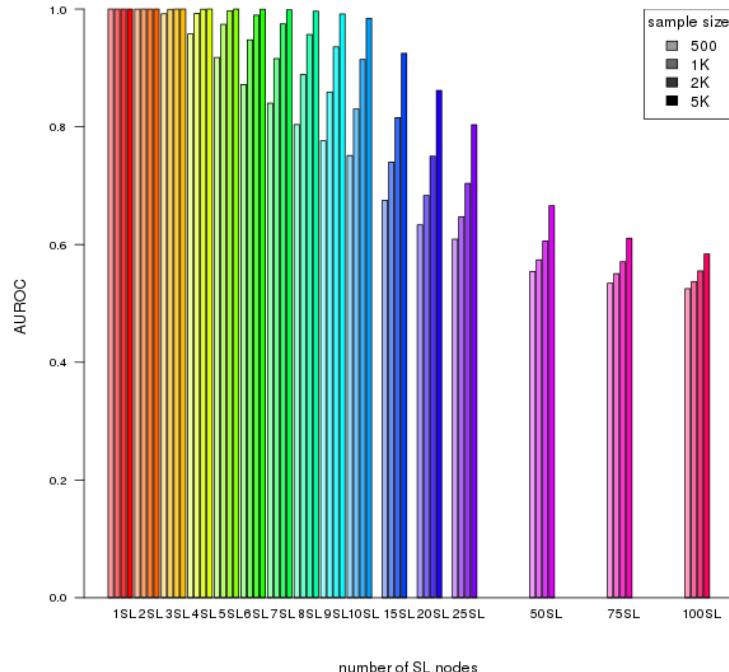
Thus SLIPT is robust across correlation structures and applicable to real gene expression data where pathway structures and correlations are a consideration. These correlation structures are not intended to model specific biological pathways or represent them, rather they serve to test the impact of correlation structure on the performance of SLIPT with an extreme example of closely correlated ($r = 0.8$) gene blocks. More complex correlation structures, such as genes positively correlated with the query gene and derived from pathway graph structures (as described in 3.4.2) will be examined below (in Section 3.3.2.2) and in Chapter 6 respectively.

In particular, genes correlated with true synthetic lethal genes have little impact on the performance of SLIPT detection: synthetic lethal genes are as distinguishable from true negative genes as without correlated genes. Synthetic lethal correlated genes will not interfere detect of true synthetic lethals, although they may be ranked next below them and be biologically informative with related gene functions.



(a) Statistical evaluation

(b) Receiver operating characteristic



(c) Statistical performance

Figure 3.17: Performance with query correlations. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (with correlation structure including correlated genes with non-SL and query genes). As before, performance of SLIPT declines for more synthetic partners and is mitigated by increased sample sizes (darker colours) but the sensitivity remains higher for a greater number of true positives with corresponding improvements in ROC curves.

3.3.2.2 Specificity with Query-Correlated Pathways

Another consideration for correlation structures is positively correlated genes with the query that are not synthetic lethal. As described in Section 3.3.2.1, 5 highly correlated ($r = 0.8$) with the query gene were added. These simulations perform similarly to before (in Figure 3.17) with a higher specificity and a lower false discovery rate being feasible (as shown in Figure 3.17a).

3.3.2.3 Importance of Directional Testing

It is important to notice here that the directional criteria of the SLIPT procedure is enhancing it's performance, particularly in distinguishing positively correlated true negatives. The multivariate normal simulation results, with 20,000 genes including all of the correlation structures discussed (SL, non-SL, and query correlated genes), are compared here for SLIPT with and without (χ^2) directional testing. There is a marked improvement in statistical performance with directional criteria, particularly with increased sensitivity and lower false discovery rate (as shown in Figure 3.18).

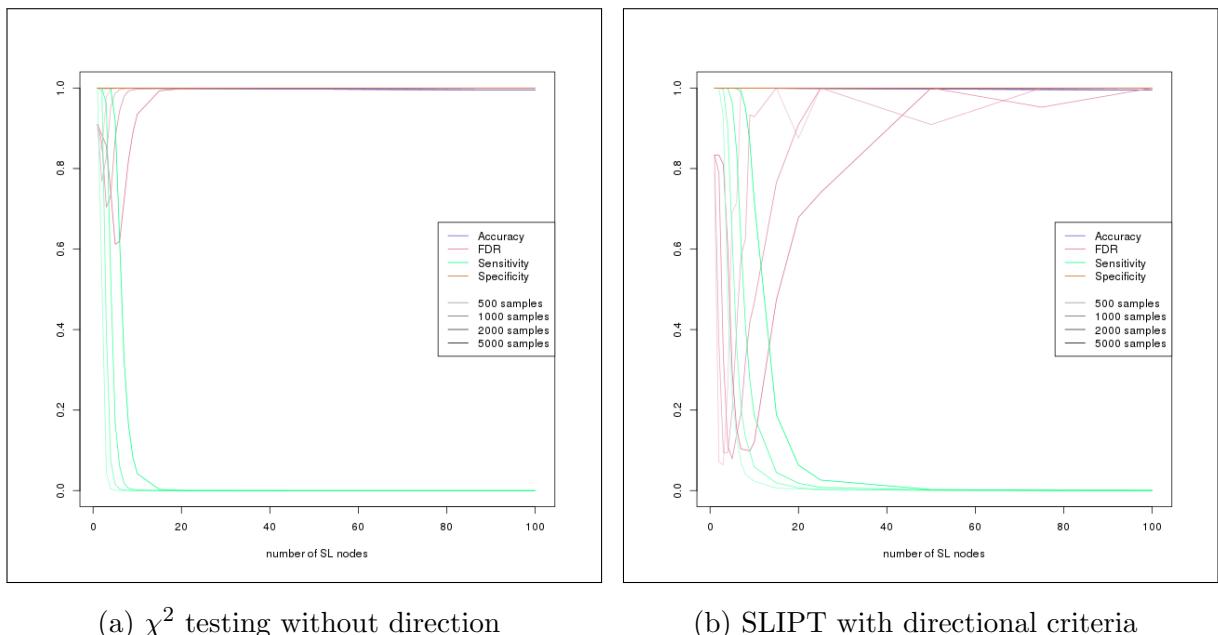
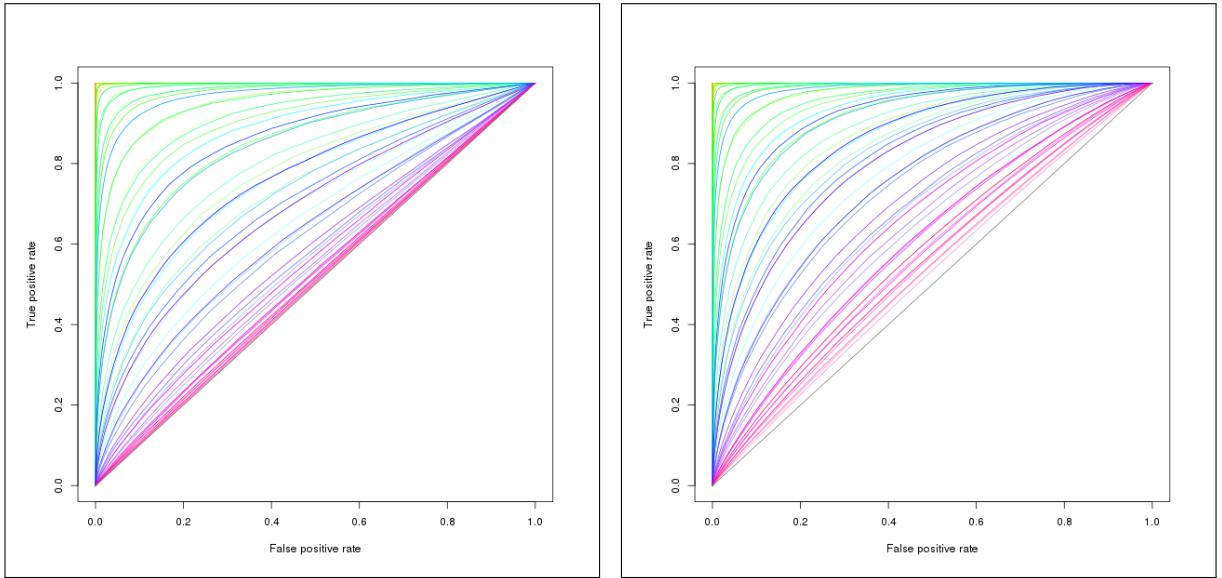
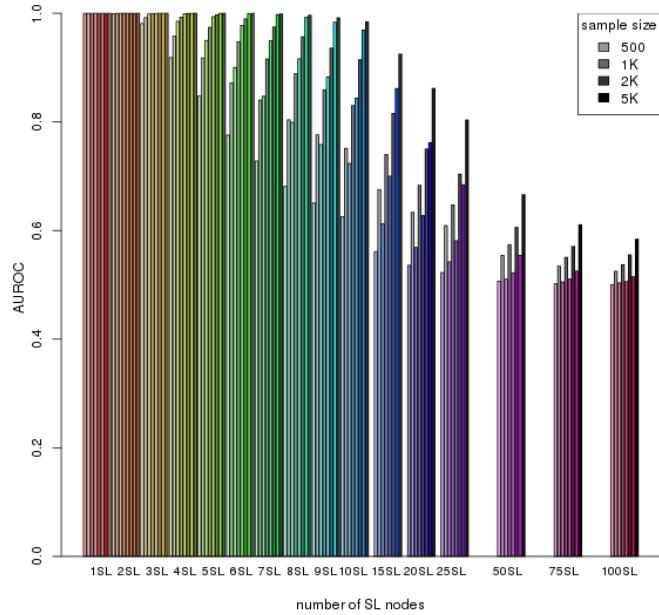


Figure 3.18: **Statistical evaluation of directional criteria.** A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and the equivalent χ^2 test without. SLIPT exhibited a consistently higher sensitivity and lower false discovery rate.



(a) χ^2 testing without direction

(b) SLIPT with directional criteria



(c) Statistical performance

Figure 3.19: Performance with directional criteria. A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and the equivalent χ^2 test without. SLIPT has higher performance across simulation parameters, clearly differing from random (grey diagonal) in ROC curves up to 100 SL genes (b). The performance (c) of SLIPT (in greyscale) was consistently higher than the χ^2 test (in color).

This is encouraging for the application of SLIPT to empirical expression datasets as positively correlated genes are likely to occur and the directional condition robustly improves the performance of SLIPT across simulation parameters. Without assuming the underlying number of synthetic lethal genes, SLIPT will perform better than the χ^2 test alone at detecting them. This is further supported irrespective of significance threshold for the χ^2 test by the ROC analysis in Figure 3.19. The directional SLIPT methodology outperforms the ordinary χ^2 test at detecting synthetic lethal partners with some predictive power (above random and AUROC of 0.5) even up to 100 synthetic lethal genes.

Together these simulation results support the application of the SLIPT methodology as it has been performed throughout Chapter 4 and 5. However, the methodology and simulation procedure will explored in more detail in Chapter 6, with the inclusion of graph structures and comparison to other synthetic lethal detection approaches.

3.4 Graph Structure Methods

Graph structures have been used in several ways in this project with novel approaches to analysis and simulations. Procedures were developed for statistical and network analysis of gene states in pathway structures. Specifically, the relationships between siRNA and SLIPT genes were tested within biological pathways in Chapter 5. These graph structures were also used in Chapter 6 for the simulation of synthetic lethality to derive correlation structure between simulated gene expression profiles in manner that resembles biological pathways.

3.4.1 Upstream and Downstream Gene Detection

Comparison of experimental and computational candidate synthetic lethal partner genes within pathway structures arose from the hypothesis that these sets of genes were related by pathway structure. Due to differences in how these candidates were generated, it should not be expected that they detect the identical genes within the candidate biological pathways, rather they may be related by being upstream or downstream of each other.

Using the Reactome version 52 data (Croft *et al.*, 2014) as described in Section 2.4.2, genes identified by each synthetic lethal discovery approach were mapped to the graph structure for the candidate pathways identified in Chapter 4 (with subgraphs defined as described in Section 2.4.3). To test whether siRNA candidate genes were upstream of SLIPT candidate genes, shortest paths were traced between each potential pair of

these genes in a directed network. The number of genes where the siRNA candidate was upstream were scored “up” and where the siRNA candidate was downstream were scored “down”. This procedure enabled counting the total number of shortest paths which supported siRNA genes being upstream or downstream of the SLIPT genes and measuring the difference between these to determine if there is an imbalance in a particular direction. While this difference is indicative of the number of paths between the gene candidate groups in either direction, alone it is not sufficient to statistically support structure or relationships between siRNA and SLIPT genes. However, it may be combined with a permutation resampling procedure (as described in Section 3.4.1.1) to test for directional relationships in either direction.

The original version of this procedure excluded gene detected by both approaches since they would count in both directions. Upon further consideration, the intersection genes were restored to being accounted for by the shortest paths counts since they may count unequally to being upstream or downstream of each gene set if there are unequal numbers above or below them in the pathway structure.

3.4.1.1 Permutation Analysis for Statistical Significance

A permutation procedure was developed to randomly assign members of the pathway to siRNA and/or SLIPT groups, with the same number of each candidate partner gene set as observed in the pathway. These permuted genes are measured for pathway structure between the permuted gene groups as performed for the observed candidates (as performed in Section 3.4.1). A distribution of pathway structure relationships expected by chance is generated by permuting iteratively over these pathways. This null distribution can be compared to the observed counts of relationships (in either direction), which yields a permutation p-value as the proportion of permutations in which had value or greater or more extreme magnitude than the observed value.

The null hypothesis is that there is no relationship between these gene groups that would not have occurred had the genes been selected at random. Thus we can test both the alternate hypothesis that the siRNA genes were upstream of the SLIPT genes or that they are downstream of them.

The permutation procedure does not assume the underlying distribution of the data under the null hypothesis and accounts for the total number of nodes, edges, siRNA, and SLIPT genes in each pathway network structure. The intersection size of the siRNA and SLIPT genes was originally not accounted for under the shortest path counts procedure that excluded them. A refined version of this procedure ensured that

the number of intersecting genes was equal to the number observed to test for pathway structure without changing the intersection size, the subject of prior analyses.

3.4.1.2 Hierarchy Based on Biological Context

An alternative approach to pathway structure was performed based on the biological context that genes at the upstream and downstream ends of a pathway perform different functions, such as a kinase signalling cascade receiving signals from external stimuli and passes these on ribosomes or the nucleus. The genes were assigned to a hierarchy to determine if genes of either candidate group (or those with stronger support for either group) performed upstream or downstream functions disproportionately.

A network-based approach was used to determine the pathway hierarchy of genes in a computationally rational way when applied to different biological pathways with a directed graph structure, G (without loops). The diameter of the network (i.e., the length of longest possible shortest path between the most distant genes) was used to identify a gene (z) at the downstream end of the pathway (at the end of a diameter spanning shortest path), assigned a hierarchy of:

$$\text{hierarchy}(z) = 1 + \text{diameter}(G)$$

Having identified the downstream end of the pathway, genes upstream (e.g., gene i) of this are assigned a hierarchy by the length of their shortest path (d) to this gene, z .

$$\text{hierarchy}(i) = \text{hierarchy}(z) - d_{iz}$$

The remaining unassigned genes (e.g., gene j) gain the hierarchy of the length of the shortest path downstream from the nearest assigned gene if possible.

$$\text{hierarchy}(j) = \text{hierarchy}(i) + d_{ij}$$

This process may be performed iteratively to fill in pathway hierarchy but it was not necessary to perform further iterations for the candidate synthetic lethal pathways investigated (amenable to this procedure) which exhibited clear directional structure and the small world property (with a low diameter). Thus genes in a pathway graph structure were assigned integer valued hierarchy upstream to downstream by this procedure:

$$\text{hierarchy} \in \{1, 2, 3, \dots, 1 + \text{diameter}(G)\}$$

This hierarchy of pathway directionality (e.g., that shown in Figure 5.7) can be used for comparison with measures of the number of genes of each candidate group and the support for being synthetic lethal partners with either approach.

3.4.2 Simulating Gene Expression from Graph Structures

A further refinement of the simulation procedure generated expression data with correlation structure, derived from a known graph structure. This enables modelling of synthetic lethal partners within a biological pathway and the investigation of impact of pathway structure on synthetic lethal prediction. A simulated pathway is first constructed as a graph structure, with the `igraph` R package Csardi and Nepusz (2006), with the added annotation of the state of the edges (i.e, whether they activate or inhibit downstream pathway members). This simulation procedure was intended for biological pathway members with correlated gene expression (higher than the background of genes in other pathways) but it may also be applicable to modelling protein levels (in a kinase regulation cascade) or substrates and products (in a metabolic pathway).

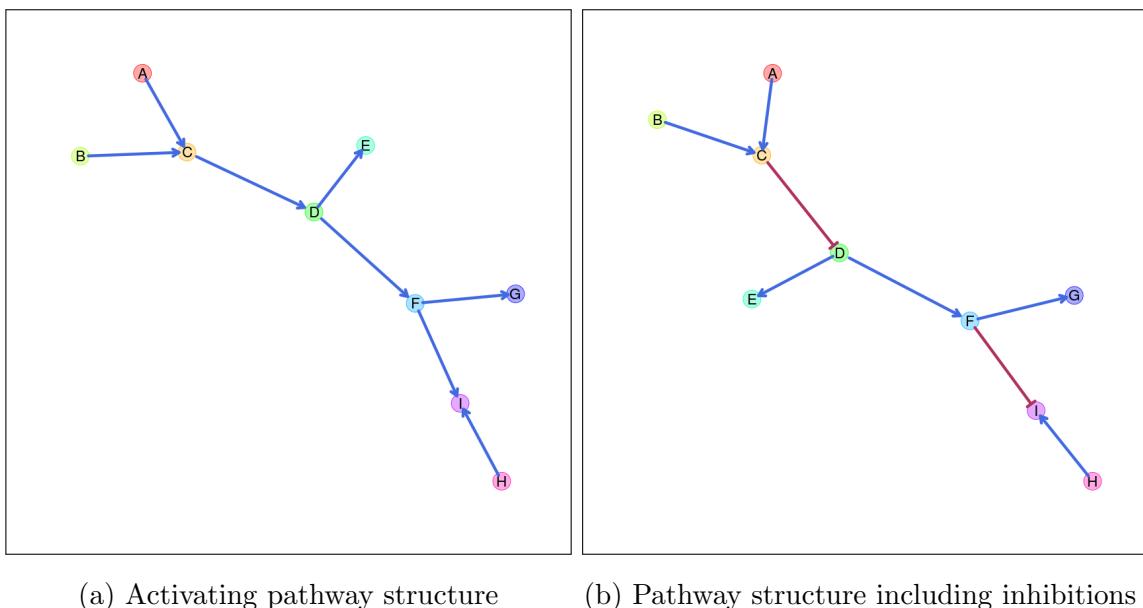


Figure 3.20: **Simulated graph structures.** A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

First, the graph structure is constructed for simulated data to be generated from (by sampling from a multivariate normal distribution using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016)). Throughout this section, the simulation procedure will be demonstrated with the relatively simple constructed graph structure shown in Figure 3.20. This graph structure visualisation was specifically developed for (directed) iGraph objects in R and has been released in the `plot.igraph` package and

`igraph.extensions` library (see Table 2.6 and Section 3.5.3). The `plot_directed` function allows customisation of plot parameters for each node or edge and mixed (directed) edge types for indicating activation or inhibition. These inhibition links (which often occur in biological pathways) are demonstrated in Figure 3.20b.

The simulation procedure is designed to use such graph structures to inform development of a “Sigma” variance-covariance matrix (Σ) for sampling from a multivariate normal distribution (using the `mvtnorm` R package). Given a graph structure (or adjacency matrix), such as Figure 3.21a, a relation matrix is calculated based on distance such that nearer nodes are given higher weight than farther nodes. For the purposes of this thesis a geometrically decreasing (relative) distance weighting is used, with each more distant node being related by $1/2$ compared to the next nearest as shown in Figure 3.21b. However, an arithmetically decreasing (absolute) distance weighting is also available in the `graphsim` R package release of this procedure.

A Σ matrix is derived from this distance weighting matrix, creating a matrix (with a diagonal of 1) where each node has a variance and standard deviation of 1. Thus covariances between adjacent nodes are assigned by a correlation parameter and the remaining matrix based on weighting these correlations with by the distance matrix (or the nearest “positive definite” matrix). For the purposes of this thesis, the correlation parameter is 0.8 unless otherwise specified (as used for the example in Figure 3.21c). This Σ matrix is used to sample from a multivariate normal distribution with each gene having a mean of 0, standard deviation 1, and covariance within the range [0, 1] such that they are correlations. This procedure generates a simulated (continuous normally distributed) expression profile for each node (as shown in Figure 3.21e) with corresponding correlation structure (Figure 3.21d). The simulated correlation structure closely resembles the expected correlation structure (Sigma in 3.21c) even for the relatively modest sample size ($N = 100$) illustrated in 3.21. Once a simulated gene expression dataset has been generated (as in Figure 3.21e), then a discrete matrix of gene function can be constructed with a functional threshold quantile to simulate functional relationships of synthetic lethality (as shown in Figure 3.4). For the purposes of this thesis, this threshold is the 0.3 quantile (as discussed in Section 3.2.1) which generates functional discrete matrices such as those used for synthetic lethal simulation in Section 3.2.2 (as shown Figure 3.21f)

The simulation procedure (depicted in Figure 3.21) is amenable to pathways containing inhibition links (as shown in Figure 3.22) with several refinements. With the inhibition links (as shown in Figure 3.22a), distances are calculated in the same manner

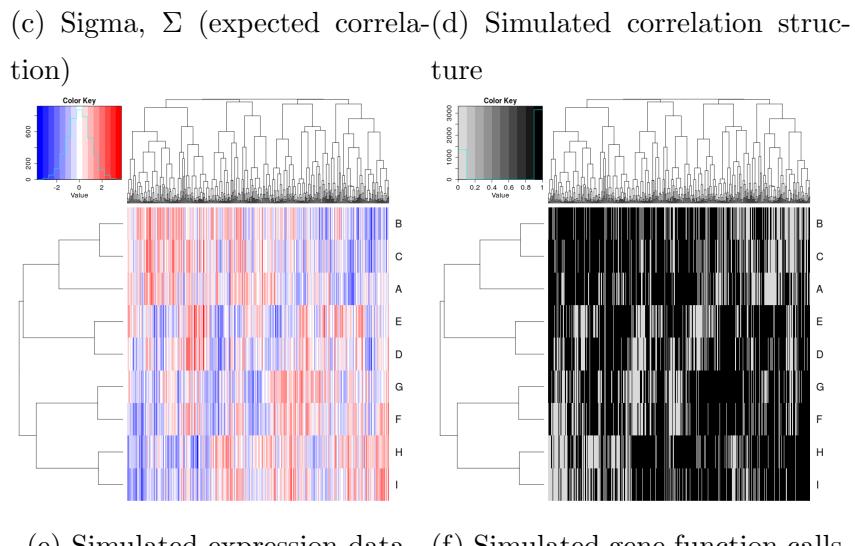
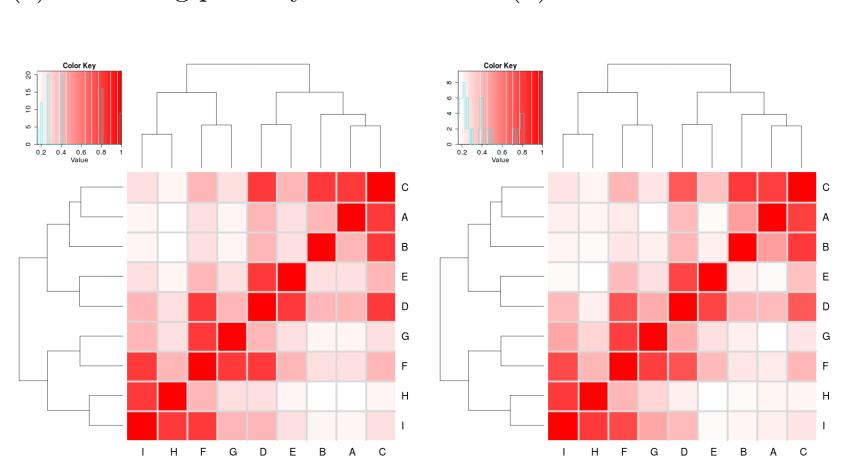
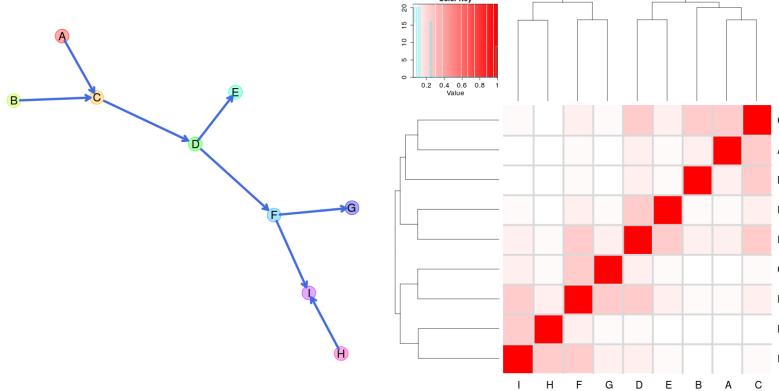


Figure 3.21: Simulating expression from a graph structure. An example graph structure is used to derive a correlation structure from the relative distances between nodes and simulate continuous gene expression with sampling from the multivariate normal distribution.

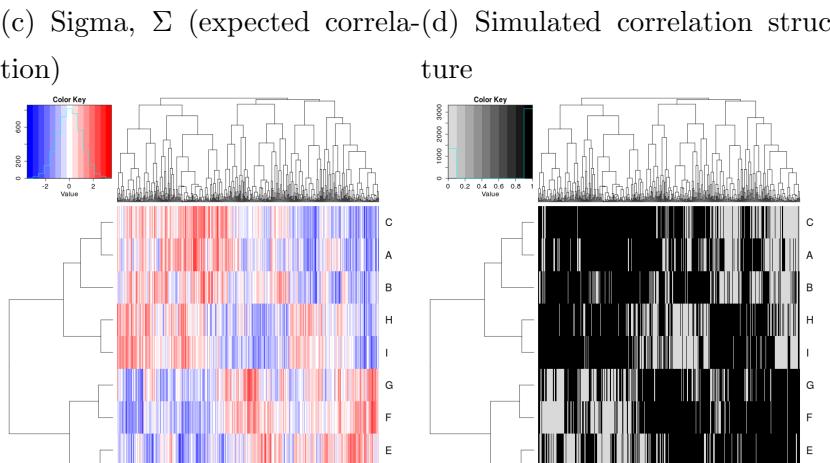
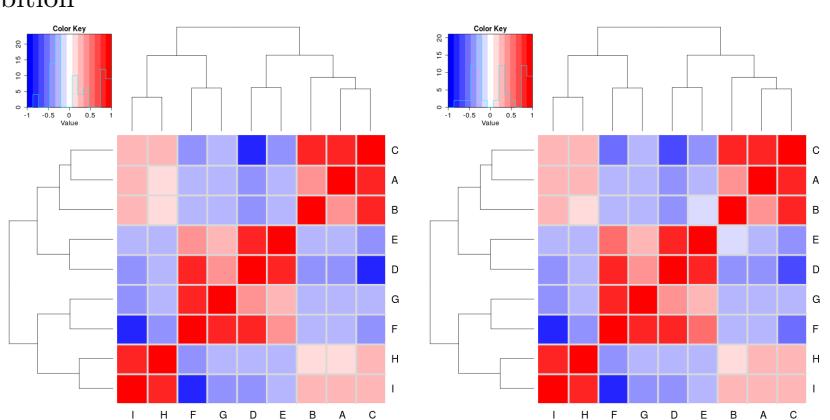
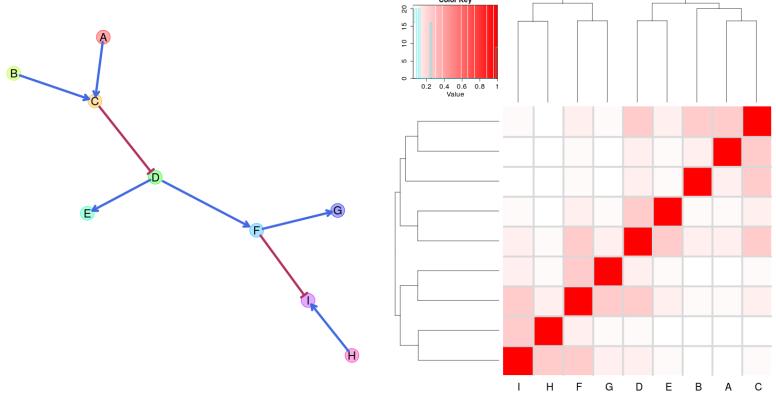


Figure 3.22: Simulating expression from graph structure with inhibitions. An example graph structure is used to derive a correlation structure from the relative distances between nodes and simulate continuous gene expression with sampling from the multivariate normal distribution.

as before (Figure 3.22b) with inhibitions accounted for by iteratively multiplying downstream nodes by -1 to form blocks of negative correlations (as shown in Figures 3.22c and 3.22d). As before, a multivariate normal distribution with these negative correlations can be sampled to generate simulated data (as shown in Figures 3.22e and 3.22f).

These simulated datasets are amenable to simulating synthetic lethal partners of a query gene within a graph network. The query gene is assumed to be separate from the graph network pathway and is added to the dataset using the procedure in Section 3.2.2. Thus we can simulate known synthetic lethal partner genes within a synthetic lethal partner pathway structure.

3.5 Customised Functions and Packages Developed

[Move to Appendix?]

Various R packages have been developed throughout this thesis using `devtools` (Wickham and Chang, 2016) and `roxygen` (Wickham *et al.*, 2017) to enable reproducibility of customised analysis and visualisation. Many of these have the added benefit of the functions being documented, demonstrated in example vignettes, and released on GitHub to enable the research community to access utilise them in their own analysis. These are summarised in Table 2.6, along with the corresponding urls for their GitHub repository which contains a README file with instructions for installation with the `devtools` R package (Wickham and Chang, 2016) and links to the relevant vignette(s) where available.

3.5.1 Synthetic Lethal Interaction Prediction Tool

The statistical methodology for detection of synthetic lethality in gene expression data (SLIPT) is one of the main novel procedures developed in this thesis, as described in Section 3.1. The `slipt` R package has been prepared for release to accompany a publication demonstrating the applications of the methodology for identifying candidate interacting genes and pathways with *CDH1* in breast cancer (TCGA, 2012).

SLIPT is amenable to analysis of any effectively continuous measure of gene activity (e.g., microarray, RNA-Seq, protein abundance, or pathway metagenes). Executing `slipt` is straightforward: the `prep_data_for_SL` function scores samples as “low”, “medium”, or “high” for each gene, then the `detect_SL` function tests a given query gene against all potential partners by performing the chi-squared test and directional conditions. This function returns a table summarising the observed and expected sample numbers used for the directional criteria, the χ^2 values, and corresponding p-values

including adjusting for multiple comparisons. The `count_of_SL` and `table_of_SL` functions serve to facilitate summary and extraction of the positive SLIPT hits, respectively, from the table of predictions of synthetic lethal partners.

The SLIPT methodology in this package release has been used in later analyses rather than the corresponding source R code, including use on remote machines and upon simulated data. In particular, the functions in the package facilitate alterations to parameters, such as the proportion of samples called as exhibiting low or high gene activity. This release support reproducibility and enables wider use of SLIPT in future investigations into other disease genes.

3.5.2 Data Visualisation

Customisations to existing data visualisations in R have been developed to present data throughout this thesis. The `vioplotx` and `heatmap.2x` packages are enhancements of the `vioplot` package (Adler, 2005) and `heatmap.2` provided by the `gplots` package (Warnes *et al.*, 2015).

The `vioplotx` package provides an alternative visualisation (of continuous variables against categories) to the more familiar boxplot, showing variability of the data by the width of the plots. As demonstrated in Figure 3.23, the customised version enables separate plotting parameters for each violin with vector inputs for colour, shape, and size of various elements of the median point, central boxplot, borders, and fill colour for the violin. Scaling violin width to adjust violin area and splitting data by a second categorical variable is also enabled. This function is intended to be backwards compatible with the inputs of `vioplot` (applying scalar inputs across all violins) and `boxplot` (by enabling formula inputs as an S3 method). Each of these features is demonstrated with examples in respective vignettes on the package GitHub repository.

The `heatmap.2x` provides extensions for annotation colour bars for both the rows and columns (as shown in Figure 3.24). Multiple bars are enabled on both axes with matrix inputs (rather than single vector for `heatmap.2`) which facilitates additional plotting of gene and sample characteristics for comparison with correlation matrices, expression profiles, or pathway metagenes. Annotation bar inputs correspond to their orientation on the plot, each colour bar is provided as a column for the row annotation on the left of the heatmap and as a row for the column annotation on top of the heatmap. Row and column annotation bars are labelled with the column or row names respectively. Additional parameters enable resizing of these annotation bar labels and control of reordering columns for if samples are ordered in advance (e.g., ranked by a

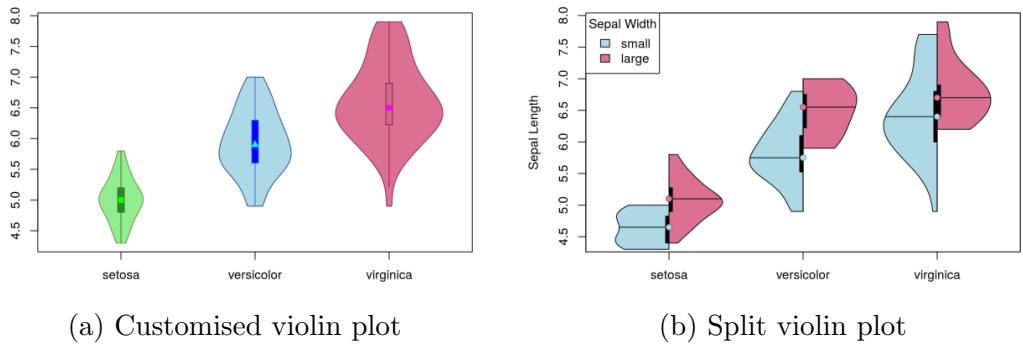


Figure 3.23: **Demonstration of violin plots with custom features.** An example of the *iris* dataset is plotted to show the custom features of the `vioplotx` package including a) individual colour, shape and size parameters of each violin, scaling violin widths by area, and b) splitting violins to compare subsets of data.

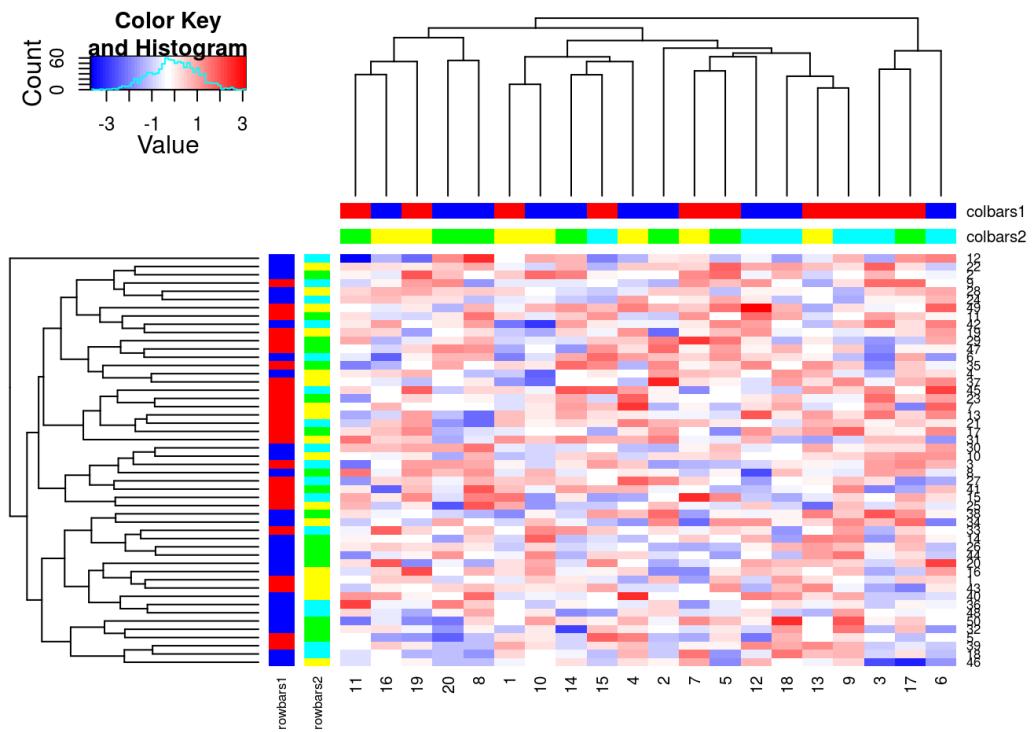


Figure 3.24: **Demonstration of annotated heatmap.** The example heatmap depicts the additional row and column annotation bars enabled by `heatmap.2x`, extending the features of `gplots` with backwards compatible inputs.

metagene or split into groups clustered separately). These features were used through this thesis and are provided in a package GitHub repository.

3.5.3 Extensions to the iGraph Package

The following features were developed during this thesis using “iGraph” data objects, building upon the `igraph` package (Csardi and Nepusz, 2006). These have been released as separate packages for each respective procedure and can be installed together as a collection of extensions to the `igraph` package.

3.5.3.1 Sampling Simulated Data from Graph Structures

The `graphsim` package implements the procedure for simulating gene expression from graph structures (as described in Section 3.4.2). By default, this derives a matrix with a geometrically decreasing weighting by distance (by shortest paths) between each pair of nodes with. An absolute decreasing weighting is also available with the option of to derive correlation structures from adjacency matrices or the number of links common partners (i.e., size of the shared “neighbourhood” (Hell, 1976)) between each pair of nodes. Functions to compute these are called directly by passing parameters to them when running the `generate_expression` or `make_sigma_mat` commands. This package enables simulating expression data directly from a graph structure (with the intermediate steps automated) or generating Σ parameters for `mvtnorm` from graph structures or matrices derived from them. These functions support assigning activating or inhibiting to each edge (with a `state` parameter).

3.5.3.2 Plotting Directed Graph Structures

The `plot.igraph` package provides the `plot_directed` function specifically developed for directed graph structures and to plot activating or inhibiting for each edge (as described in Section 3.4.2). As shown in Figure 3.25, this function supports separate plotting parameters for each node, node label, and edge. This includes colours of node fill, border, label text, and edges and size of nodes, edge widths, arrowhead lengths, and font size of labels. The `state` parameter for assigning activating or inhibiting to each edge determines whether edges are depicted with 30° or 90° arrowheads. Colours are assigned separately and so they may be customised. Vectorised parameters are applied across each node or edge, whereas scalar parameters apply the same plotting parameters across them. The default layout function is `layout.fruchterman.reingold` but any layout function supported by `plot` function in `igraph` (Csardi and Nepusz, 2006) is compatible such as `layout.kamada.kawai` used to implement the Kamada–Kawai algorithm (Kamada and Kawai, 1989) for graph plots throughout this thesis.

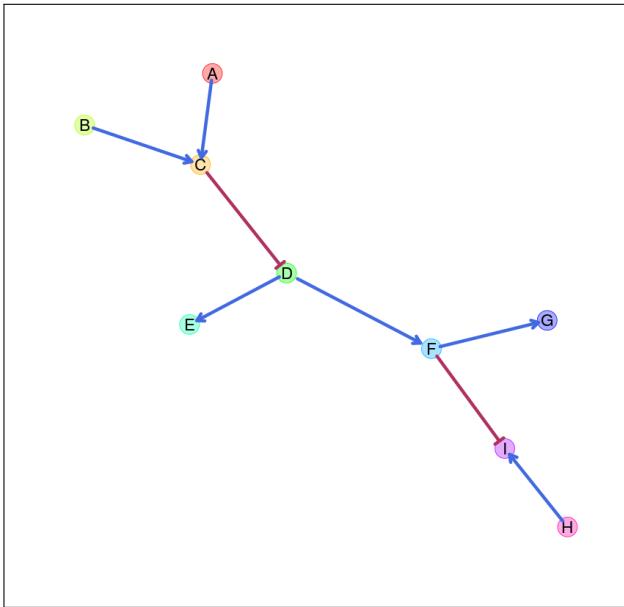


Figure 3.25: **Simulating graph structures.** An example graph structure which will be used throughout demonstrating the simulation procedure from graph structures. Here activating links are denoted by blue arrows and inhibiting links by red edges.

3.5.3.3 Computing Information Centrality

The shortest paths of a network are computed by the `igraph` package Csardi and Nepusz (2006) which can be extended to calculate the network efficiency but is not provided by the package itself (as described in Section 2.4.4). The “information centrality” of a vertex is computed as the relative change in the network efficiency when the vertex is removed. Information centrality is calculated iteratively for each node and the sum of information centrality for each vertex is the information centrality for the network. These metrics are provided by the `info.centrality` package.

3.5.3.4 Testing Pathway Structure with Permutation Testing

A network-based procedure developed was used for comparison of siRNA and SLIPT candidate genes in a pathway structure. Such pathway structure relationships were tested by computing the number of shortest paths between two different groups of nodes in either direction within a graph. This pathway relationship metric was implemented in the `pathway.structure.permutation` package with permutation testing (as described in sections 3.4.1 and 3.4.1.1).

3.5.3.5 Metapackage to Install iGraph Functions

These features may be installed together with `igraph.extensions` which can be accessed from a GitHub repository. This meta-package installs `igraph` (Csardi and Nepusz, 2006) and the packages described in Section 3.5.3 including their dependencies for matrix operations and statistical procedures: `Matrix`, `matrixcalc`, and `mvtnorm` (Bates and Maechler, 2016; Genz and Bretz, 2009; Genz *et al.*, 2016; Novomestky, 2012).

Chapter 4

Synthetic Lethal Analysis of Gene Expression Data

Having developed a statistical synthetic lethal detection methodology, SLIPT, it was then applied to publicly available cancer gene expression datasets. The analysis focuses on breast cancer for which TCGA expression data (TCGA, 2012) from a patient cohort and siRNA screen data (Telford *et al.*, 2015) from experiments conducted in MCF10A cells were available. Stomach cancer data (Bass *et al.*, 2014) was used to replicate findings in an independent dataset, with this cancer chosen because it also occurs in syndromic HDGC patients. The TCGA data also has the advantages of other clinical and molecular profiles including somatic mutation across many of the same samples, in addition to a considerable sample size for RNASeq expression data generated with a common TCGA procedures to minimise batch effects.

Synthetic lethal candidate partners for *CDH1* were identified at both the gene and pathway level. SLIPT gene candidates were analysed by cluster analysis for common expression profiles across samples and relationships with clinical factors and mutations in key breast cancer genes. These genes will also be compared to the gene candidates from a primary and secondary (validation) screens conducted by Telford *et al.* (2015) on isogenic cell lines. For comparison, the SLIPT methodology was also applied using mutation data for *CDH1* against expression of candidate partners (as described in Section 3.1) which may better represent the null mutations in HDGC patients and the experiment cell model (Chen *et al.*, 2014). Pathways were analysed by over-representation analysis (with resampling for comparisons with siRNA data) and supported by a meta-gene analysis of pathway gene signatures. The pathway metagene expression profiles were used to replicate known relationships between clinical and molecular characteris-

tics for breast cancer and to demonstrate application of SLIPT directly on metagenes to detect synthetic lethal pathways.

Together these results demonstrate the wide range of applications for SLIPT analysis and examine the synthetic lethal partners of *CDH1* in breast and stomach cancer. These synthetic lethal genes and pathways were identified both in the context of the functional implications of novel synthetic lethal relationships and as potential actionable targets against *CDH1* deficient tumours, in addition to replication of established functions of E-cadherin. In particular, these analyses focused on comparisons with experimental screening data to explore the potential for SLIPT to augment triage of candidate partners and support further experimental investigations. The key synthetic lethal partner pathways for *CDH1*, supported by both approaches, will be examined in more detail at the gene and pathway structure level in Chapter 5.

4.1 Synthetic Lethal Genes in Breast Cancer

The SLIPT methodology (as described in Section 3.1) was applied to the normalised TCGA breast cancer gene expression dataset ($n = 1168$). As shown in Table 4.1, the most significant genes had strong evidence of expression-based association with *CDH1* (high χ^2 values) with fewer samples exhibiting low expression of both genes than expected statistically. Eukaryotic translation genes were among the highest scoring gene candidates, including initiation factors, elongation factors, and ribosomal proteins. These are clearly necessary for cancer cells to grow and proliferate, with sustained gene expression needed to maintain growth signalling pathways and resist apoptosis or immune factors translation may be subject to non-oncogene addiction for *CDH1*-deficient cells.

While these are among the strongest synthetic lethal candidates, translational genes are crucial to the viability of healthy cells and dosing for a selective synthetic lethal effect against these may be difficult compared to other biological functions which may also be supported among the SLIPT candidate genes. Furthermore, few known biological functions of *CDH1* were among the strongest SL candidates, so the remaining candidate genes may also be informative since they are likely to contain these expected functions in addition to novel relationships for *CDH1*. Thus further pathway level analyses were also conducted to examine biological functions over-represented among synthetic candidate genes and to identify synthetic lethal pathways.

The modified mtSLIPT methodology (as described in Section 3.1) was also applied to the normalised TCGA breast cancer gene expression dataset, against somatic loss

Table 4.1: Candidate synthetic lethal gene partners of *CDH1* from SLIPT

Gene	Observed	Expected	χ^2 value	p-value	p-value (FDR)
<i>TRIP10</i>	62	130	162	5.65×10^{-34}	1.84×10^{-31}
<i>EEF1B2</i>	56	130	158	3.10×10^{-33}	9.45×10^{-31}
<i>GBGT1</i>	61	131	156	1.08×10^{-32}	3.14×10^{-30}
<i>ELN</i>	81	130	149	3.46×10^{-31}	8.82×10^{-29}
<i>TSPAN4</i>	78	130	146	1.63×10^{-30}	3.79×10^{-28}
<i>GLIPR2</i>	72	130	146	1.68×10^{-30}	3.86×10^{-28}
<i>RPS20</i>	73	131	145	1.89×10^{-30}	4.28×10^{-28}
<i>RPS27A</i>	80	130	143	5.53×10^{-30}	1.18×10^{-27}
<i>EEF1A1P9</i>	63	130	141	1.91×10^{-29}	3.74×10^{-27}
<i>C1R</i>	73	130	141	2.05×10^{-29}	3.97×10^{-27}
<i>LYL1</i>	73	130	140	2.99×10^{-29}	5.74×10^{-27}
<i>RPLP2</i>	71	130	139	4.88×10^{-29}	9.07×10^{-27}
<i>C10orf10</i>	73	130	138	6.72×10^{-29}	1.20×10^{-26}
<i>DULLARD</i>	74	131	138	9.29×10^{-29}	1.61×10^{-26}
<i>PPM1F</i>	64	130	136	1.61×10^{-28}	2.65×10^{-26}
<i>OBFC2A</i>	69	130	136	2.49×10^{-28}	3.93×10^{-26}
<i>RPL11</i>	70	130	136	2.56×10^{-28}	3.97×10^{-26}
<i>RPL18A</i>	70	130	135	3.08×10^{-28}	4.70×10^{-26}
<i>MFNG</i>	76	131	133	7.73×10^{-28}	1.12×10^{-25}
<i>RPS17</i>	77	131	133	8.94×10^{-28}	1.29×10^{-25}
<i>MGAT1</i>	73	130	132	1.44×10^{-27}	2.03×10^{-25}
<i>RPS12</i>	72	130	128	8.57×10^{-27}	1.12×10^{-24}
<i>C10orf54</i>	73	130	127	1.37×10^{-26}	1.75×10^{-24}
<i>LOC286367</i>	72	130	126	2.20×10^{-26}	2.70×10^{-24}
<i>GMFG</i>	70	130	126	2.20×10^{-26}	2.70×10^{-24}

Strongest candidate SL partners for *CDH1* by SLIPT with observed and expected numbers of TCGA breast cancer samples with low expression of both genes.

of function mutations in *CDH1*. As shown in Table C.1, the most significant genes also had strong evidence of expression associated with *CDH1* mutations (high χ^2 values) with fewer samples with *CDH1* exhibiting low expression each candidate gene than expected statistically. These genes were not strongly supported as the expression analysis (in Table 4.1), however, nor were as many genes detected. This is perhaps unsurprising due to the lower sample size with matching somatic mutation data and the lower frequency of *CDH1* mutations compared to low expression defined by 1/3 quantiles.

The mtSLIPT candidates had more genes involved in cell and gene regulation, particularly DNA and RNA binding factors. The strongest candidates also included

microtubule (*KIF12*), microfibril (*MFAP4*), and cell adhesion (*TENC1*) genes consistent with the established cytoskeletal role of *CDH1*. The elastin gene (*ELN*) was notably strongly supported by both expression and mutation SLIPT analysis of *CDH1* supporting interactions with extracellular proteins and the tumour microenvironment.

4.1.1 Synthetic Lethal Pathways in Breast Cancer

Translational pathways were strongly over-represented in SLIPT partners, as shown in Table 4.2. These include ribosomal subunits, initiation, peptide elongation, and termination. Regulatory processes involving mRNA including 3' untranslated region (UTR) binding, L13a-mediated translational silencing, and nonsense-mediated decay were also implicated. These are consistent with protein translation being subject to “non-oncogene addiction” (Luo *et al.*, 2009), as a core process that is dysregulated to sustain cancer proliferation and survival (Gao and Roux, 2015).

Immune pathways, including the adaptive immune system and responses to infectious diseases were also strongly implicated as synthetic lethal with loss of E-cadherin. This is consistent with the alterations of immune response being a hallmark of cancer Hanahan and Weinberg (2000), since evading the immune system is necessary for cancer survival. Either of these systems are potential means to target *CDH1* deficient cells, although these were not detected in an isolated cell line experimental screen (Telford *et al.*, 2015) and the differences between the findings in patient data are described in more detail in Section 4.2.5.

It is also notable that the pathways over-represented in SLIPT candidate genes have strongly significant over-representation of Reactome pathways from the hypergeometric test (as described in Section 2.3.2). Even after adjusting stringently for multiple tests, biologically related pathways were supported together. These pathways are further supported by testing for synthetic lethality against *CDH1* mutations (mtSLIPT) with many of these pathways also among the most strongly supported in this analysis (shown in Table C.2). This mutation-based analysis more closely represents the null *CDH1* mutations in HDGC (Guilford *et al.*, 1998) and the experimental MCF10A cell model (Chen *et al.*, 2014). There was still support for translational and immune pathways not detected in the isolated experimental system. G-protein-coupled receptors (GPCRs) also among the most strongly supported pathways, supporting the experimental findings of Telford *et al.* (2015) for these intracellular signalling pathways already being targeted for other diseases.

Table 4.2: Pathways for *CDH1* partners from SLIPT

Pathways Over-represented	Pathway Size	SL Genes	p-value (FDR)
Eukaryotic Translation Elongation	86	81	1.3×10^{-207}
Peptide chain elongation	83	78	5.6×10^{-201}
Eukaryotic Translation Termination	83	77	1.2×10^{-196}
Viral mRNA Translation	81	76	1.2×10^{-196}
Formation of a pool of free 40S subunits	93	81	3.7×10^{-194}
Nonsense Mediated Decay independent of the Exon Junction Complex	88	77	5.3×10^{-187}
L13a-mediated translational silencing of Ceruloplasmin expression	103	82	9.6×10^{-183}
3' -UTR-mediated translational regulation	103	82	9.6×10^{-183}
GTP hydrolysis and joining of the 60S ribosomal subunit	104	82	1.9×10^{-181}
Nonsense-Mediated Decay	103	80	6.2×10^{-176}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	103	80	6.2×10^{-176}
Adaptive Immune System	412	167	6.5×10^{-174}
Eukaryotic Translation Initiation	111	82	5.7×10^{-173}
Cap-dependent Translation Initiation	111	82	5.7×10^{-173}
SRP-dependent cotranslational protein targeting to membrane	104	79	2.0×10^{-171}
Translation	141	91	6.1×10^{-170}
Infectious disease	347	146	1.6×10^{-166}
Influenza Infection	117	81	1.9×10^{-163}
Influenza Viral RNA Transcription and Replication	108	77	1.9×10^{-160}
Influenza Life Cycle	112	77	2.5×10^{-156}

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for *CDH1*.

4.1.2 Expression Profiles of Synthetic Lethal Partners

Due to the sheer number of gene candidates, investigations proceeded into correlation structure and pathway over-representation. These analyses also examined expression patterns of synthetic lethal gene candidates. This serves to explore the functional similarity of the synthetic lethal partners of *CDH1*, with the eventual aim to assess their utility as drug targets. As shown in Figure 4.1 (which clusters *CDH1* lowly expressing samples separately), there were several large clusters of genes among the expression profiles of the *CDH1* synthetic lethal candidate partners. The clustering suggests co-regulation of genes or pathway correlation between partner gene candidates. A number of candidates from an experimental RNAi screen study performed by Telford *et al.* (2015) were also identified by this approach. In addition, we identified novel gene candidates, which had not been observed affect viability in isogenic cell line experiments.

In these expression profiles, a gene with a moderate or high signal across samples exhibiting low *CDH1* expression would represent a potential drug target. However, it appears that several molecular subtypes of cancer have elevation of different clusters of synthetic lethal candidates in samples with low *CDH1*. This clustering suggests

that different targets (or combinations) could be effective in different patients, suggesting potential utility for stratification. In particular, estrogen receptor negative, basal-like subtype, and “normal-like” tumours (Dai *et al.*, 2015; Eroles *et al.*, 2012; Parker *et al.*, 2009) have elevation of genes specific to particular clusters, indicative of some synthetic lethal interactions being specific to a particular molecular subtype or genetic background. Thus synthetic lethal drug therapy against these subtypes may be ineffective if it were designed against genes in another cluster.

A similar correlation structure was observed among the candidates tested against *CDH1* mutation (mtSLIPT), as shown in Figure C.1. This clustering analysis similarly identified several major clusters of putative synthetic lethal partner genes. In this case, many partner genes had consistently high expression across most of the (predominantly lobular subtype) *CDH1* breast cancer samples. However, a major exception to this in the *CDH1* expression analysis were the normal samples which were excluded from the mutation data (as they were not tested for tumour-specific genotypes). This supports synthetic lethal interventions being more applicable to *CDH1* mutant tumours. There was still considerable correlation structure, particularly among *CDH1* wildtype samples, sufficient to distinguish gene clusters. In contrast to the expression analysis the (predominantly ductal *CDH1* wildtype) basal-like subtype and estrogen receptor negative samples had depleted expression among most candidate synthetic lethal partners. This is consistent with synthetic lethal interventions only being effective in lobular estrogen receptor positive breast cancers in which they are a more common, as recurrent (driver) mutation. However, the remaining samples are still informative for synthetic lethal analysis (by SLIPT) as it requires highly expressing *CDH1* samples for comparison.

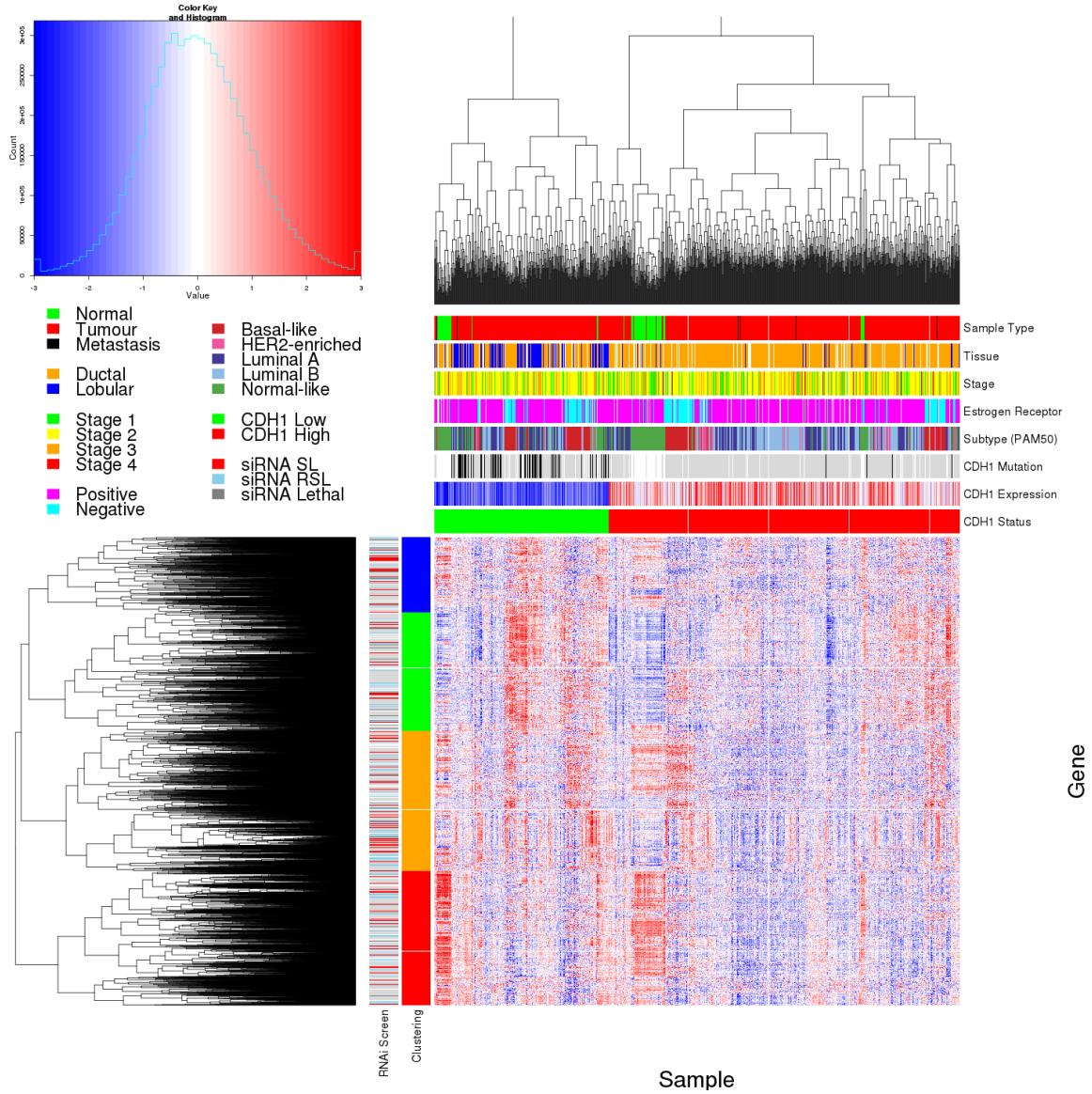


Figure 4.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance, complete linkage) of all samples (separated by the 1/3 quantile of *CDH1* expression) analysed in TCGA breast cancer dataset for gene expression of 5,165 candidate partners of E-cadherin (*CDH1*) from SLIPT prediction (with FDR adjusted $p < 0.05$). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or lethal in an siRNA screen (Telford *et al.*, 2015). Screen results for synthetic lethal (SL), the reverse effect (RSL), or lethal cell viability are shown as reported by Telford *et al.* (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in *CDH1* low samples, notably ‘normal-like’, ‘basal-like’, and estrogen receptor negative samples have elevated expression in one or more distinct clusters showing complexity and variation among candidate synthetic lethal partners. *CDH1* low samples also contained most of samples with *CDH1* mutations (shown in black). Negative values for mutation and screen data are shown in light grey with missing data in white.

The *CDH1* mutant samples (in Figure 4.1) were predominantly among the low *CDH1* expressing samples and distributed throughout *CDH1* samples with clustering analysis. Thus the molecular profiles of *CDH1* low samples were indistinguishable from *CDH1* mutant samples, with the exception of normal samples (that do not have somatic mutation data available). Conversely, many of the *CDH1* mutant samples (in Figure C.1) had among the lowest *CDH1* expression, and some of the synthetic lethal partners were also highly expressed in low expressing *CDH1* wildtype samples, despite these not being considered as “inactivated” by mtSLIPT analysis.

Together these results support the use of low *CDH1* expression as a strategy for detecting *CDH1* inactivation. This has the benefit of increasing sample size (including samples such as normal tissue which do not have somatic mutation data available) and increasing the expected number of mutually inactive (low-low) samples for the directional criteria of (mt)SLIPT which enabling it to better distinguish significant deviations below this (as discussed in Section 6.1). This also circumvents the assumption that all (detected) mutations are inactivating (although synonymous mutations were excluded from the analysis), which may not be the case for several highly expressing *CDH1* mutant samples that do not cluster together in Figures 4.1 or C.1. One of these exhibits among the lowest expression for many predicted synthetic lethal partners and would not be vulnerable to inactivation of these genes. As such correctly genotyping inactivating mutations will be essential in clinical practice for synthetic lethal targeting tumour suppressor genes, particularly for other genes such as *TP53* where oncogenic and tumour suppressor mutations (with different molecular consequences) are both common in cancers. Using expression as a measure of gene expression also avoids the assumptions that mutations are somatic rather than germline and that gene inactivation is by detectable mutations rather than other mechanisms such as epigenetic changes which is supported by many lowly expressing *CDH1* wildtype samples clustering with similar profiles to mutant samples.

4.1.2.1 Subgroup Pathway Analysis

Synthetic lethal gene candidates for *CDH1* from SLIPT analysis of RNA-Seq gene expression data were also used for pathway over-representation analyses (as described in Section 2.3.2). The correlation structure in the expression of candidates synthetic lethal genes in *CDH1* low tumours (lowest 1/3rd quantile of expression) was examined for distinct biological pathways in subgroups of genes elevated in different clusters of samples. These genes were highly expressed in different samples with their clini-

Table 4.3: Pathway composition for clusters of *CDH1* partners from SLIPT

Pathways Over-represented in Cluster 1	Pathway Size	Cluster Genes	p-value (FDR)
Collagen formation	67	10	4.0×10^{-11}
Extracellular matrix organisation	238	21	1.8×10^{-9}
Collagen biosynthesis and modifying enzymes	56	8	1.8×10^{-9}
Uptake and actions of bacterial toxins	22	5	9.5×10^{-9}
Elastic fibre formation	37	6	1.9×10^{-8}
Muscle contraction	62	7	2.4×10^{-7}
Fatty acid, triacylglycerol, and ketone body metabolism	117	10	4.9×10^{-7}
XBP1(S) activates chaperone genes	51	6	6.6×10^{-7}
IRE1alpha activates chaperones	54	6	1.2×10^{-6}
Neurotoxicity of clostridium toxins	10	3	1.3×10^{-6}
Retrograde neurotrophin signalling	10	3	1.3×10^{-6}
Assembly of collagen fibrils and other multimeric structures	40	5	1.9×10^{-6}
Collagen degradation	58	6	2.0×10^{-6}
Arachidonic acid metabolism	41	5	2.1×10^{-6}
Synthesis of PA	26	4	3.0×10^{-6}
Signalling by NOTCH	80	7	3.3×10^{-6}
Signalling to RAS	27	4	3.7×10^{-6}
Integrin cell surface interactions	82	7	4.2×10^{-6}
Pathways Over-represented in Cluster 2	Pathway Size	Cluster Genes	p-value (FDR)
Eukaryotic Translation Elongation	86	75	1.1×10^{-181}
Viral mRNA Translation	81	72	9.8×10^{-179}
Peptide chain elongation	83	72	1.9×10^{-175}
Eukaryotic Translation Termination	83	72	1.9×10^{-175}
Formation of a pool of free 40S subunits	93	75	1.9×10^{-171}
Nonsense Mediated Decay independent of the Exon Junction Complex	88	72	9.9×10^{-168}
L13a-mediated translational silencing of Ceruloplasmin expression	103	75	3.0×10^{-159}
3' -UTR-mediated translational regulation	103	75	3.0×10^{-159}
Nonsense-Mediated Decay	103	75	3.0×10^{-159}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	103	75	3.0×10^{-159}
SRP-dependent cotranslational protein targeting to membrane	104	75	3.2×10^{-158}
GTP hydrolysis and joining of the 60S ribosomal subunit	104	75	3.2×10^{-158}
Eukaryotic Translation Initiation	111	75	4.5×10^{-151}
Cap-dependent Translation Initiation	111	75	4.5×10^{-151}
Influenza Infection	117	75	1.4×10^{-145}
Influenza Viral RNA Transcription and Replication	108	72	5.7×10^{-145}
Translation	141	81	8.0×10^{-143}
Influenza Life Cycle	112	72	2.3×10^{-141}
Pathways Over-represented in Cluster 3	Pathway Size	Cluster Genes	p-value (FDR)
Adaptive Immune System	412	90	6.1×10^{-61}
Chemokine receptors bind chemokines	52	27	6.7×10^{-56}
Generation of second messenger molecules	29	21	6.5×10^{-55}
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	64	29	6.5×10^{-55}
TCR signalling	62	27	8.9×10^{-51}
Peptide ligand-binding receptors	161	40	1.5×10^{-45}
Translocation of ZAP-70 to Immunological synapse	16	14	3.1×10^{-43}
Costimulation by the CD28 family	51	22	4.0×10^{-43}
PD-1 signalling	21	15	4.0×10^{-41}
Class A/1 (Rhodopsin-like receptors)	258	50	6.7×10^{-41}
Phosphorylation of CD3 and TCR zeta chains	18	14	1.3×10^{-40}
Interferon gamma signalling	74	24	5.0×10^{-39}
GPCR ligand binding	326	57	1.8×10^{-38}
Cytokine Signalling in Immune system	268	48	8.9×10^{-37}
Downstream TCR signalling	45	18	1.8×10^{-35}
G _{αi} signalling events	167	33	2.2×10^{-33}
Cell surface interactions at the vascular wall	99	21	1.3×10^{-26}
Interferon Signalling	164	28	1.7×10^{-26}
Pathways Over-represented in Cluster 4	Pathway Size	Cluster Genes	p-value (FDR)
Extracellular matrix organisation	238	48	8.0×10^{-41}
Class A/1 (Rhodopsin-like receptors)	258	47	2.8×10^{-36}
GPCR ligand binding	326	54	2.1×10^{-34}
G _{αs} signalling events	83	22	1.4×10^{-31}
GPCR downstream signalling	472	68	1.1×10^{-29}
Haemostasis	423	61	3.3×10^{-29}
Platelet activation, signalling and aggregation	180	31	7.1×10^{-28}
Binding and Uptake of Ligands by Scavenger Receptors	40	14	9.9×10^{-27}
RA biosynthesis pathway	22	11	2.5×10^{-26}
Response to elevated platelet cytosolic Ca ²⁺	82	19	3.0×10^{-26}
Developmental Biology	420	57	3.5×10^{-26}
G _{αi} signalling events	167	28	7.3×10^{-26}
Platelet degranulation	77	18	1.6×10^{-25}
Gastrin-CREB signalling pathway via PKC and MAPK	171	28	2.5×10^{-25}
Muscle contraction	62	16	4.7×10^{-25}
G _{αo} signalling events	150	25	3.2×10^{-24}
Retinoid metabolism and transport	34	12	5.0×10^{-24}
Phase 1 - Functionalisation of compounds	67	16	6.5×10^{-24}

Pathway over-representation analysis for Reactome pathways with the number of genes in each pathway (Pathway Size), number of genes within the pathway identified (Cluster Genes), and the pathway over-representation p-value (adjusted by FDR) from the hypergeometric test.

cal factors including estrogen receptor status and intrinsic subtype, from the PAM50 procedure (Parker *et al.*, 2009) shown in Figure 4.1.

As shown by the most over-represented pathways in Table 4.3, each correlated cluster of candidate synthetic lethal partners of *CDH1* contains functionally different genes. Cluster 1 contains genes with less evidence of over-represented pathways than other clusters, corresponding to less correlation between genes within the cluster, and to it being a relatively small group. While there is some indication that collagen biosynthesis, microfibril elastic fibres, extracellular matrix, and metabolic pathways may be over-represented in Cluster 1, these results are mainly based on small pathways containing few synthetic lethal genes. Genes in Cluster 2 exhibited low expression in normal tissue samples compared to tumour samples (see Figure 4.1) and show compelling evidence of over-representation of post-transcriptional gene regulation and protein translation processes. Similarly, Cluster 3 has over-representation of immune signalling pathways (including chemokines, secondary messenger, and TCR signalling) and downstream intracellular signalling cascades such as G protein coupled receptor (GPCR) and G_{αi} signalling events. While pathway over-representation was weaker among genes in Cluster 4, they contained intracellular signalling pathways and were highly expressed in normal samples (in contrast to Cluster 2). Cluster 4 also involved extracellular factors and stimuli such as extracellular matrix, platelet activation, ligand receptors, and retinoic acid signalling.

Based on these results, potential synthetic lethal partners of *CDH1* include processes known to be dysregulated in cancer, such as translational, cytoskeletal, and immune processes. Intracellular signalling cascades such as the GPCRs and extracellular stimuli for these pathways were also implicated in potential synthetic lethality with *CDH1*.

Similar translational, cytoskeletal, and immune processes were identified among SLIPT partners with respect to *CDH1* mutation, shown in Table C.3. While GPCR signalling was replicated in mtSLIPT analysis, there was also stronger over-representation for NOTCH, ERBB2, and PI3K/AKT signalling in mutation analysis consistent with these signals being important for proliferation of *CDH1* deficient tumours. The GPCR and PI3K/AKT pathways are of particular interest as pathways with oncogenic mutations that can be targeted and downstream effects on translation (a strongly supported process across analyses). Extracellular matrix pathways (e.g., elastic fibre formation) were also supported across analyses (in Tables 4.3 and C.3) consistent with the estab-

lished cell-cell signalling role of *CDH1* and the importance of the tumour microenvironment for cancer proliferation.

4.2 Comparing Synthetic Lethal Gene Candidates

4.2.1 Primary siRNA Screen Candidates

Gene candidates were compared between computational (SLIPT in TCGA breast cancer data) and experimental (the primary siRNA screen performed by Telford *et al.* (2015)) approaches in Figure 4.2. The number of genes detected by both methods did not produce a significant overlap but these may be difficult to compare due to vast differences between the detection methods. There were similar issues in the comparison of mtSLIPT genes tested against *CDH1* mutations (in Appendix Figure ??), despite excluding genes not tested by both methods in either test. However, these intersecting genes may still be functionally informative or amenable to drug triage as they were replicated across both methods and pathway over-representation differed between the sections of the Venn diagram (see Figure 4.2).

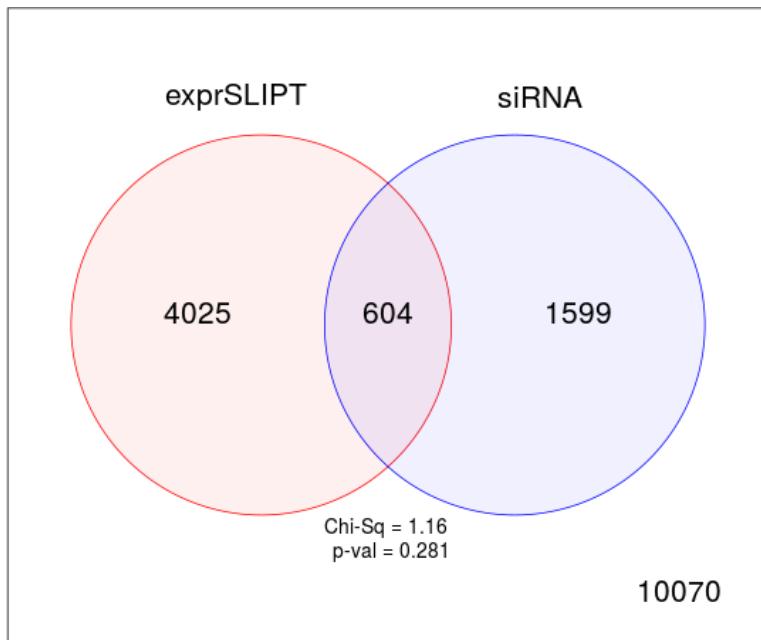


Figure 4.2: **Comparison of SLIPT to siRNA.** Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ^2 test suggests that the overlap is no more than would be expected by chance ($p = 0.281$). Only genes tested by both methods were included.

4.2.2 Comparison with Correlation

Another potential means to triage drug target candidates is by correlation of expression profiles with *CDH1*. Correlation with *CDH1* was compared to SLIPT and siRNA results in Figure 4.3. The genes not detected by SLIPT (including siRNA candidates) had included gene with insignificant SLIPT p-values. As expected, these genes were distributed around a correlation of zero and genes with higher correlation with *CDH1* (either direction) were more significant, although there were exceptions to this trend and larger positive correlations than negative correlations. The majority of SLIPT candidates had negative correlations, particularly genes detected by both approaches, although these were typically weak correlations and are unlikely to be sufficient to detect such genes on their own. This is supported by simulation results in Section 6.1.

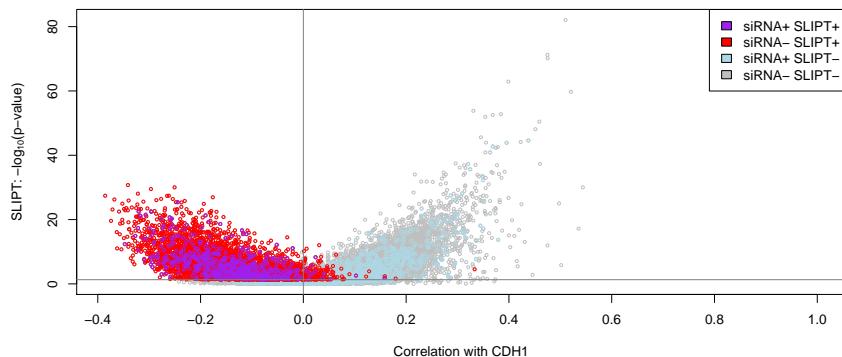


Figure 4.3: Compare SLIPT and siRNA genes with correlation. The χ^2 p-values for genes tested by SLIPT (in TCGA breast cancer) expression analysis were compared against Pearson’s correlation of gene expression with *CDH1*. Genes detected by SLIPT or siRNA are coloured according to the legend.

There were not strong positive correlations with *CDH1* among siRNA candidates, consistent with previous findings that co-expression is not predictive of synthetic lethality (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015). Negative correlation may not be indicative of synthetic lethality either as many siRNA candidates also had positive correlations. The SLIPT methodology has shown to detect genes with both positive and negative correlations, although it does appear to preferentially detect negatively correlated genes to some extent. These findings were replicated with the mtSLIPT approach against *CDH1* mutation (in Figure C.3), although the range of the χ^2 p-values differ due to lower sample size for mutation analysis.

The apparent tendency for genes detected by SLIPT or siRNA to have negative correlations with *CDH1* expression is not due to the smaller number of genes in these groups. The distribution of *CDH1* correlations differed across these gene groups (as shown by Figures 4.4 and C.4), specifically lower in SLIPT candidates (as supported by ANOVA in Table 4.4). However, these are relatively weak correlations and further triage of gene candidates by correlation is not suitable, nor is use of correlation itself to predict synthetic lethal partners in the first place.

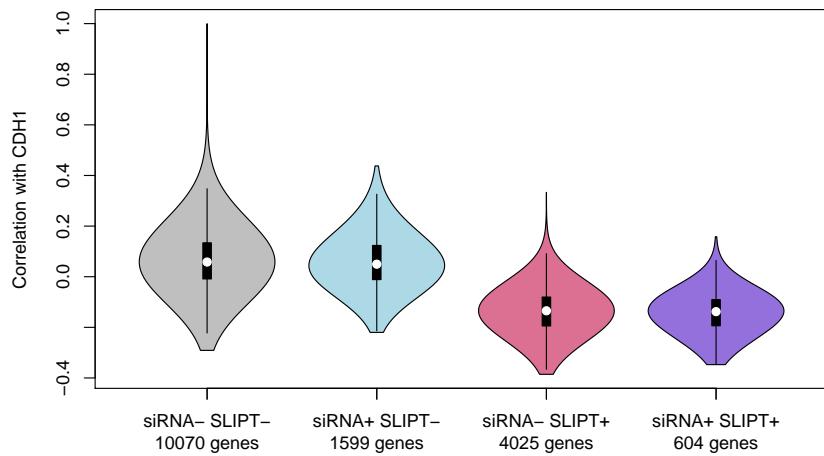


Figure 4.4: Compare SLIPT and siRNA genes with correlation. Genes detected as candidate synthetic lethal partners by SLIPT (in TCGA breast cancer) expression analysis and experimental screening (with siRNA) were compared against Pearson's correlation of gene expression with *CDH1*. There were no differences in correlation between gene groups detected by either approach.

Table 4.4: ANOVA for Synthetic Lethality and Correlation with *CDH1*

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.027	0.027	2.8209	0.09306
SLIPT	1	134.603	134.603	14115.9824	<0.0001
siRNA×SLIPT	1	0.000	0.000	0.0073	0.93212

Analysis of variance for correlation with *CDH1* against synthetic lethal detection approaches (with an interaction term). Only genes tested by both methods were included in this analysis.

4.2.3 Comparison with Primary Screen Viability

A similar comparison of SLIPT results was made with the viability ratio (of *CDH1* mutant to wildtype) in the primary siRNA screen performed by Telford *et al.* (2015). The significance and viability thresholds used for SLIPT and siRNA detection of synthetic lethal candidate partners of *CDH1* are shown in Figure 4.5. However, not all of the genes below the viability thresholds were necessarily selected to be candidate partners, as additional criteria were used in each case: directional criteria as for SLIPT (see Section 3.1) and minimum wildtype viability for siRNA (Telford *et al.*, 2015).

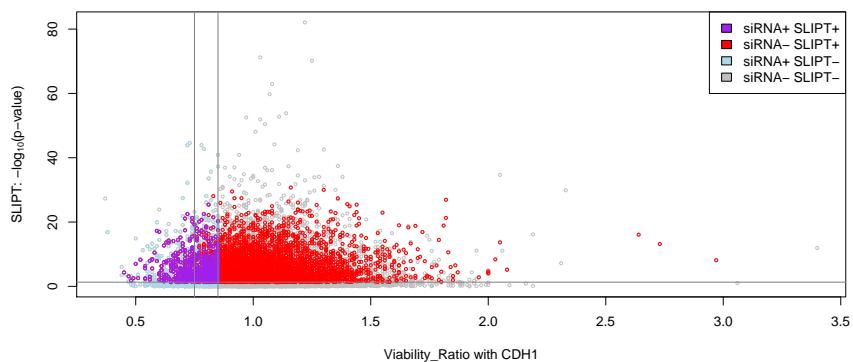


Figure 4.5: Compare SLIPT and siRNA genes with viability. The χ^2 p-values for genes tested by SLIPT (in TCGA breast cancer) expression analysis were compared (on a log-scale) against the viability ratio of *CDH1* mutant and wildtype cells in the primary siRNA screen. Genes detected by SLIPT or siRNA are coloured according to the legend.

There does not appear to be a clear relationship between SLIPT and siRNA candidates. Many genes not detected by both approaches were numerous in Figures 4.2 and C.2. These genes detected by either are not necessarily near the thresholds for the other. In this respect the SLIPT approach with patient data and cell line experiments are independent means to identify synthetic lethal candidates. While genes detected by both approaches were not necessarily more strongly supported by either, the genes with a viability closer to 1 (no synthetic lethal effect) in siRNA included those with more significant SLIPT p-values whereas more extreme viability ratios tended to be less significant (as shown by Figure 4.5). However, it should be noted that genes with more moderate viability ratios were more common and SLIPT was capable (despite adjusting for multiple testing) of detecting significant genes with extreme viability ratios, particularly those considerably lower than 1.

However, there was not little support for SLIPT candidates having considerably different viability ratios (as shown in Figures 4.6 and C.5). While the viability thresholds used by Telford *et al.* (2015) to detect synthetic lethal candidates in the primary screen, the genes identified by SLIPT had a higher mean viability ratio (by t-test: $t = 2.1553$, $p = 0.03117$). However, the effect size was small (mean SLIPT⁻ 1.029, mean SLIPT⁺ 1.037) and the vast majority of SLIPT candidate genes did not have different viability in the primary screen to genes not identified by SLIPT.

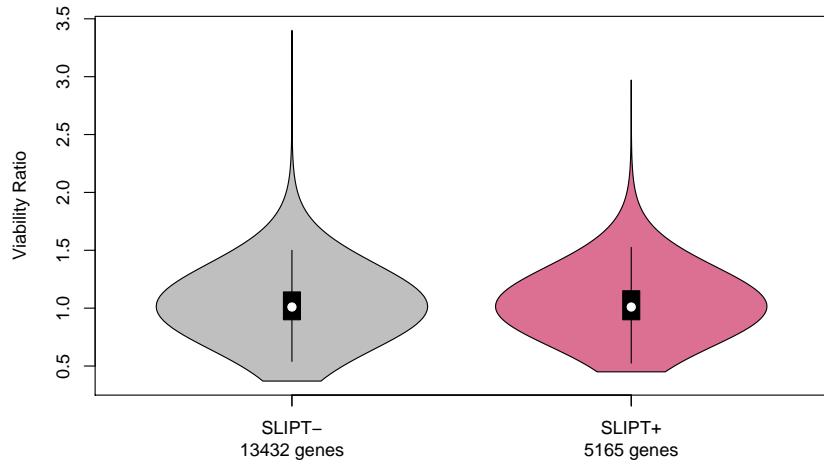


Figure 4.6: Compare SLIPT genes with siRNA viability. Genes detected as candidate synthetic lethal partners by SLIPT (in TCGA breast cancer) expression analysis were compared against the viability ratio of *CDH1* mutant and wildtype cells in the primary siRNA screen. There were clear no differences in viability between genes detected by SLIPT and those not with the differences being primarily due to viability thresholds being used to detect synthetic lethality by Telford *et al.* (2015).

4.2.4 Comparison with Secondary siRNA Screen Validation

However, it should be noted that genes with a lower viability ratio were not necessarily the most strongly supported by experimental screening. The primary screen (with 4 pooled siRNAs) has been used for the majority of comparisons in this thesis because the genome-wide panel of target genes screened enables a large number of genes to be compared with SLIPT results from gene expression and somatic mutation analysis. A secondary screen was also performed by Telford *et al.* (2015) on the isogenic MCF10A breast cell lines to validate the individual (i.e., non-pooled) siRNAs separately, with

the strongest candidates being those exhibiting synthetic lethal viability ratios replicated across independently targeting siRNAs. The strongest candidates from a primary screen were subject to a further secondary screen for validation by independent replication with 4 gene knockdowns with different targeting siRNAs. This was performed for the top 500 candidates (with the lowest viability ratio) from the primary screen and the 482 of these genes also tested by SLIPT in breast cancer.

The secondary screen results show that SLIPT candidate genes were more significantly ($p = 7.49 \times 10^{-3}$ by Fisher's exact test) more likely to be validated in the secondary screen and are thus informative of more robust partner genes, in addition to providing support that these interactions are consistent with expression profiles from heterogeneous patient samples across genetic backgrounds. As shown in Table 4.5, there is significant association between SLIPT candidates and stronger validations of siRNA candidates. Since there were more SLIPT⁻ genes among those not validated and more SLIPT⁺ genes among those validated with several siRNAs, this supports the use of SLIPT as a synthetic lethal discovery procedure which may augment such screening experiments.

Table 4.5: Comparing SLIPT genes against secondary siRNA screen in breast cancer

		Secondary Screen					Total
		0/4	1/4	2/4	3/4	4/4	
SLIPT⁺	Observed	70	46	31	8	2	157
	Expected	85	44	10	4	2	
SLIPT⁻	Observed	190	90	31	10	4	325
	Expected	175	91	42	12	4	
Total		280	136	52	18	6	482

While the individual genes detected by either approach do not necessarily match (and are potentially false-positives), the biological functions important in *CDH1* deficient cancers and potential mechanisms for specific targeting of them can be further supported by pathway analysis of the gene detected by either method. The genes detected by both approaches may therefore be more informative at the pathway level, where it is unlikely for a pathway to be consistently detected by chance. As the SLIPT candidates differ from the siRNA candidates (and are more likely to be validated), they can provide additional mechanisms by which *CDH1* deficient cancers proliferate

and vulnerabilities that may be exploited against them by using the synthetic lethal pathways.

4.2.5 Comparison to Primary Screen at Pathway Level

These pathway over-representation analyses (performed as described in Section 2.3.2) correspond to genes separated into SLIPT or siRNA screen candidates unique to either method or detected by both (Table 4.6). The SLIPT-specific gene candidates were involved most strongly with translational and immune regulatory pathways, although extracellular matrix pathways were also supported. These pathways were largely consistent with those identified in Table 4.2 and in the clustering analysis (Table 4.3). The genes detected only by the siRNA screen had over-representation of cell signalling pathways, including many containing genes known to be involved in cancer (e.g., MAPK, PDGF, ERBB2, and FGFR), with the detection of Class A GPCRs supporting the independent analyses by Telford *et al.* (2015). The intersection of computational and experimental synthetic lethal partners of *CDH1* had stronger evidence for over-representation of GPCR pathways and more specific subclasses, such as visual phototransduction ($p = 6.9 \times 10^{-10}$) and $G_{\alpha s}$ signalling events ($p = 1.7 \times 10^{-7}$), than other signalling pathways.

The pathway analysis for mtSLIPT against *CDH1* mutations (in Table C.4) had concordant results for both mtSLIPT-specific and siRNA-specific pathways. While the specific pathway composition of the intersection of these analyses differed from SLIPT against low *CDH1* expression, signalling pathways including GPCRs, NOTCH, EERB2, PDGF, and SCF-KIT. These findings indicate the signalling pathways are among the most suitable vulnerability to exploit in targeting *CDH1* deficient tumours as they can be detected in both a patient cohort (with TCGA expression data) and tested in a laboratory system. However, it is possible that the isolated experimental system is set up to preferentially detect kinase singalling pathways (which are amenable to pharmacological inhibition and translation to the clinic) and the other pathways identified by SLIPT may still be informative of the role of *CDH1* loss of function in cancers or mechanisms by which further gene loss leads to specific inviability.

Table 4.6: Pathway composition for *CDH1* partners from SLIPT and siRNA screening

Predicted only by SLIPT (4025 genes)	Pathway Size	Genes Identified	p-value (FDR)
Eukaryotic Translation Elongation	80	75	1.5×10^{-182}
Peptide chain elongation	77	72	2.9×10^{-176}
Viral mRNA Translation	75	70	4.9×10^{-172}
Eukaryotic Translation Termination	76	70	5.9×10^{-170}
Formation of a pool of free 40S subunits	87	74	9.5×10^{-166}
Nonsense Mediated Decay independent of the Exon Junction Complex	81	70	1.2×10^{-160}
L13a-mediated translational silencing of Ceruloplasmin expression	97	75	3.8×10^{-155}
3' -UTR-mediated translational regulation	97	75	3.8×10^{-155}
GTP hydrolysis and joining of the 60S ribosomal subunit	98	75	6.0×10^{-154}
Nonsense-Mediated Decay	96	73	5.2×10^{-150}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	96	73	5.2×10^{-150}
SRP-dependent cotranslational protein targeting to membrane	97	73	7.8×10^{-149}
Eukaryotic Translation Initiation	105	75	4.7×10^{-146}
Cap-dependent Translation Initiation	105	75	4.7×10^{-146}
Translation	133	83	4.0×10^{-142}
Influenza Viral RNA Transcription and Replication	102	71	2.9×10^{-137}
Influenza Infection	111	74	3.7×10^{-137}
Influenza Life Cycle	106	71	2.3×10^{-133}
Infectious disease	326	125	4.2×10^{-120}
Extracellular matrix organisation	189	77	5.4×10^{-95}

Detected only by siRNA screen (1599 genes)	Pathway Size	Genes Identified	p-value (FDR)
Class A/1 (Rhodopsin-like receptors)	282	44	1.3×10^{-27}
GPCR ligand binding	363	52	5.8×10^{-26}
G _{αq} signalling events	159	26	6.7×10^{-23}
Gastrin-CREB signalling pathway via PKC and MAPK	180	27	2.0×10^{-21}
G _{αi} signalling events	184	27	5.3×10^{-21}
Downstream signal transduction	146	23	7.6×10^{-21}
Signalling by PDGF	172	25	4.0×10^{-20}
Peptide ligand-binding receptors	175	25	8.5×10^{-20}
Signalling by ERBB2	146	22	1.3×10^{-19}
DAP12 interactions	159	23	2.6×10^{-19}
DAP12 signalling	149	22	2.7×10^{-19}
Organelle biogenesis and maintenance	264	33	5.5×10^{-19}
Signalling by NGF	266	33	8.2×10^{-19}
Downstream signalling of activated FGFR1	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR2	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR3	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR4	134	20	1.1×10^{-18}
Signalling by FGFR	146	21	1.3×10^{-18}
Signalling by FGFR1	146	21	1.3×10^{-18}
Signalling by FGFR2	146	21	1.3×10^{-18}

Intersection of SLIPT and siRNA screen (604 genes)	Pathway Size	Genes Identified	p-value (FDR)
Visual phototransduction	54	9	6.9×10^{-10}
G _{αs} signalling events	48	7	1.6×10^{-7}
Retinoid metabolism and transport	24	5	1.7×10^{-7}
Acyl chain remodelling of PS	10	3	6.5×10^{-6}
Transcriptional regulation of white adipocyte differentiation	51	6	6.5×10^{-6}
Chemokine receptors bind chemokines	22	4	6.5×10^{-6}
Signalling by NOTCH4	11	3	6.9×10^{-6}
Defective EXT2 causes exostoses 2	11	3	6.9×10^{-6}
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	11	3	6.9×10^{-6}
Platelet activation, signalling and aggregation	146	12	6.9×10^{-6}
Phase 1 - Functionalisation of compounds	41	5	1.3×10^{-5}
Amine ligand-binding receptors	13	3	1.7×10^{-5}
Acyl chain remodelling of PE	14	3	2.4×10^{-5}
Signalling by GPCR	300	23	2.4×10^{-5}
Molecules associated with elastic fibres	29	4	2.6×10^{-5}
DAP12 interactions	128	10	2.6×10^{-5}
Cytochrome P ₄₅₀ - arranged by substrate type	30	4	3.2×10^{-5}
GPCR ligand binding	147	11	3.8×10^{-5}
Acyl chain remodelling of PC	16	3	4.0×10^{-5}
Response to elevated platelet cytosolic Ca ²⁺	66	6	4.2×10^{-5}

4.2.5.1 Resampling Genes for Pathway Enrichment

Comparisons of genes between experimental screen candidates and prediction from TCGA expression data were less consistent than comparisons of pathways. However, this is not unexpected, since synthetic lethal pathways are more robustly conserved (Dixon *et al.*, 2008) and the computational approach using patient samples from complex tumour microenvironment has considerably different strengths to an experimental screen (Telford *et al.*, 2015) based on genetically homogenous cell line models in an isolated laboratory environment. For instance, it is unlikely for immune signalling to be detected in an isolated cell culture system.

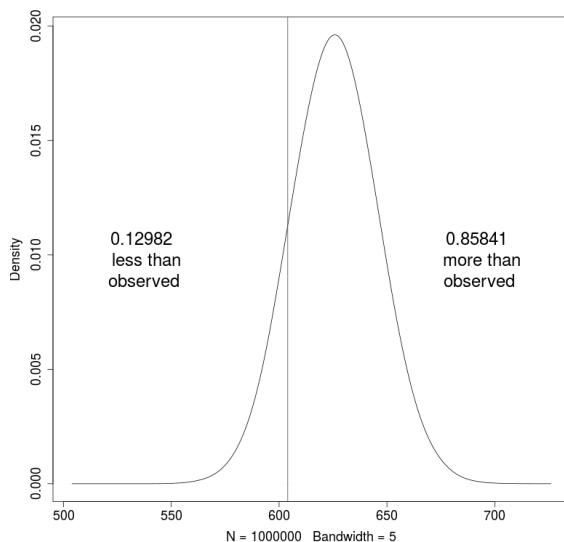


Figure 4.7: Resampled intersection of SLIPT and siRNA candidates. Resampling analysis of intersect size from genes detected by SLIPT and siRNA screening approaches over 1 million replicates. The proportion of expected intersection sizes for random samples below or above the observed intersection size respectively, lacking significant over-representation or depletion of siRNA screen candidates within the SLIPT predictions for *CDH1*.

The overlap between synthetic lethal candidates from bioinformatics SLIPT predictions and siRNA screening has raised other questions, including whether the pathways over-represented would be expected by chance. This of particular concern since the siRNA candidate genes themselves are highly over-represented for particular pathways (e.g., GPCRs) so selecting any intersect with them could be enriched for these pathways. Another pathway-based approach is to test whether pathways are over-represented in randomly sampled genes, comparing many “resamplings” or “permutations” of these genes to the enrichment statistics observed for these pathways in the

SLIPT candidates and their intersection with the siRNA hits shows whether we detect these pathways more than we expect by chance (as described in Section 2.3.6).

Of particular concern are the over-represented pathways in genes detected by both methods. Pathway over-representation alone does not detect whether SLIPT predicted genes or siRNA candidates are enriched within each other. This resampling analysis therefore detects whether over-represented pathways were detected by SLIPT independently of their over-representation among siRNA candidates (without assuming an underlying test statistic distribution).

A resampling approach is also applicable to testing whether the number of genes detected by each approach significantly intersected. As shown in Figure 4.7, resampling did not find evidence of significant depletion or over-representation for experimental synthetic lethal candidate genes in the computationally predicted synthetic lethal partners of *CDH1*, and thus the observed overlap may be due to chance. This is consistent with previous findings (see Figure 4.2) and does not preclude pathway relationships being supported by resampling.

A permutation analysis was performed to resample the genes tested by both approaches to investigate whether the observed pathway over-representation could have occurred in a randomly selected sample of genes from the experimental candidates, that is, whether the pathway predictions from SLIPT could be expected by chance (as described in Sections 2.2.3.1 and 2.3.6). While the number of siRNA candidate genes also detected by SLIPT was not statistically significant ($p = 0.281$), this may be due to the vastly different limitations of the approaches and the correlation structure of gene expression not being independent (as assumed for multiple testing procedures). The intersection may still be functionally relevant to *CDH1*-deficient cancers, such as the pathway data in Table 4.6. The resampling analysis for pathways was compared to the pathway over-representation for SLIPT predicted synthetic lethal partners in Table 4.7. Similarly, the pathway resampling for intersection between SLIPT predictions and experimental screen candidates was compared to pathway over-representation in Table 4.8 for intersection with siRNA data.

The pathway resampling approach for SLIPT-specific gene candidates (Table 4.7) replicates the gene set over-representation analysis for all SLIPT genes, detecting evidence of synthetic lethal pathways for *CDH1* in translational, immune, and cell signalling pathways including G_{αi} signalling, GPCR downstream signalling, and chemokine receptor binding. While the immune and signal transduction pathways were not significantly over-represented in the resampling analysis, the results for the two approaches

were largely consistent for translation and post-transcriptional gene regulation, supporting gene set over-representation of the SLIPT-specific pathways in Table 4.7. In particular, some of the most significantly over-represented pathways had higher observed χ^2 values than any of the 1 million random permutations. Similar pathways were also replicated by permutation analysis for mtSLIPT candidate partners against *CDH1* mutation (shown in Table C.5). This shows that many of the pathways detected specifically by SLIPT are replicated by permutation procedures and that the permutation approach is capable of detecting many of the most strongly over-represented pathways.

Table 4.7: Pathways for *CDH1* partners from SLIPT

Reactome Pathway	Over-representation	Permutation
Eukaryotic Translation Elongation	1.3×10^{-207}	$< 1.241 \times 10^{-5}$
Peptide chain elongation	5.6×10^{-201}	$< 1.241 \times 10^{-5}$
Viral mRNA Translation	1.2×10^{-196}	$< 1.241 \times 10^{-5}$
Eukaryotic Translation Termination	1.2×10^{-196}	$< 1.241 \times 10^{-5}$
Formation of a pool of free 40S subunits	3.7×10^{-194}	$< 1.241 \times 10^{-5}$
Nonsense Mediated Decay independent of the Exon Junction Complex	5.3×10^{-187}	$< 1.241 \times 10^{-5}$
L13a-mediated translational silencing of Ceruloplasmin expression	9.6×10^{-183}	$< 1.241 \times 10^{-5}$
3' -UTR-mediated translational regulation	9.6×10^{-183}	$< 1.241 \times 10^{-5}$
GTP hydrolysis and joining of the 60S ribosomal subunit	1.9×10^{-181}	$< 1.241 \times 10^{-5}$
Nonsense-Mediated Decay	6.2×10^{-176}	$< 1.241 \times 10^{-5}$
Nonsense Mediated Decay enhanced by the Exon Junction Complex	6.2×10^{-176}	$< 1.241 \times 10^{-5}$
Adaptive Immune System	6.5×10^{-174}	0.15753
Eukaryotic Translation Initiation	5.7×10^{-173}	$< 1.241 \times 10^{-5}$
Cap-dependent Translation Initiation	5.7×10^{-173}	$< 1.241 \times 10^{-5}$
SRP-dependent cotranslational protein targeting to membrane	2.0×10^{-171}	$< 1.241 \times 10^{-5}$
Translation	6.1×10^{-170}	$< 1.241 \times 10^{-5}$
Infectious disease	1.6×10^{-166}	0.23231
Influenza Infection	1.9×10^{-163}	$< 1.241 \times 10^{-5}$
Influenza Viral RNA Transcription and Replication	1.9×10^{-160}	$< 1.241 \times 10^{-5}$
Influenza Life Cycle	2.5×10^{-156}	$< 1.241 \times 10^{-5}$
<i>Extracellular matrix organisation</i>	1.1×10^{-152}	0.071761
GPCR ligand binding	1.1×10^{-143}	0.55801
Class A/1 (Rhodopsin-like receptors)	1.5×10^{-142}	0.58901
<i>GPCR downstream signalling</i>	7.6×10^{-140}	0.098357
Haemostasis	1.9×10^{-134}	0.27059
Developmental Biology	2.0×10^{-123}	0.52737
Metabolism of lipids and lipoproteins	3.3×10^{-120}	0.724
Cytokine Signalling in Immune system	2.6×10^{-119}	0.39661
Peptide ligand-binding receptors	3.7×10^{-109}	0.61102
<i>G_{αi} signalling events</i>	8.9×10^{-100}	$< 1.241 \times 10^{-5}$

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

The permutation approach was then also applied to the intersection between computational and experimental candidates. The permutation analysis is testing for consistent detection of pathways was independent of their pre-existing status as experimental candidates. The pathway results for these candidate partners (in Table 4.8) differed between over-representation and resampling analyses.

Namely, many of the over-represented pathways were not significant in the resampling analysis, including visual phototransduction and retinoic acid signalling, and were likely over-represented in the intersection due to over-representation in the siRNA candidates rather than additional support from SLIPT. In contrast, pathways involving defective *EXT1* or *EXT2* genes approach significance after FDR adjustment for multiple tests in resampling. Of the highest over-represented pathways in the intersection, only $G_{\alpha s}$ signalling events were supported by both over-representation and resampling analyses. Other pathways supported by both analyses were cytoplasmic elastic fibre formation, associated HS-GAG protein modification pathways, energy metabolism, and the fibrin clotting cascade.

Many of the pathways supported in the intersection by permutation analysis were also replicated in the mtSLIPT analysis of partners tested with *CDH1* mutation (in Table C.6), including $G_{\alpha s}$, elastic fibres, HS-GAG, and energy metabolism. While there were differences between the pathways identified by over-representation analysis, those replicated by permutation were highly concordant, supporting the combined use of these pathway approaches to identify synthetic lethal gene functions and targets.

While this indicates that $G_{\alpha s}$ and GPCR class A/1 signalling events were significantly detected by both approaches, GPCR signalling pathways overall were not. It is likely that GPCRs were primarily over-represented in the intersection with the experimental candidates due to strong over-representation of these pathways in experimental candidates, rather than detection by SLIPT, which may be driven by these more specific constituent pathways.

However, several pathways, including some immune functions and neurotransmitters, were supported by the resampling analysis (in Tables 4.8 and C.6) when the initial pathway over-representation test was not significant. These functions appear to have been detected by both approaches more than expected by chance but must be interpreted with caution since they were still not common enough to be detected in pathway over-representation analysis.

4.2.6 Integrating Synthetic Lethal Pathways and Screens

Based on these results, it appears that computational and experimental approaches to synthetic lethal screening for *CDH1* lead to a broader functional characterisation, and many candidate partners, when combined, despite different strengths and limitations. Compared to candidate gene approaches, experimental genome-wide screens are an appealing unbiased strategy for identifying synthetic lethal interactions. Since these

Table 4.8: Pathways for *CDH1* partners from SLIPT and siRNA primary screen

Reactome Pathway	Over-representation	Permutation
Visual phototransduction	6.9×10^{-10}	0.91116
G_{as} signalling events	1.6×10^{-7}	0.012988
Retinoid metabolism and transport	1.7×10^{-7}	0.20487
Transcriptional regulation of white adipocyte differentiation	6.5×10^{-6}	0.38197
Acylic chain remodelling of PS	6.5×10^{-6}	0.58485
Chemokine receptors bind chemokines	6.5×10^{-6}	0.97255
<i>Defective EXT2 causes exostoses 2</i>	6.9×10^{-6}	0.056437
<i>Defective EXT1 causes exostoses 1, TRPS2 and CHDS</i>	6.9×10^{-6}	0.056437
Signalling by NOTCH4	6.9×10^{-6}	0.15497
Platelet activation, signalling and aggregation	6.9×10^{-6}	0.53358
Phase 1 - Functionalisation of compounds	1.3×10^{-5}	0.24836
Amine ligand-binding receptors	1.7×10^{-5}	0.3195
Acylic chain remodelling of PE	2.4×10^{-5}	0.7307
Signalling by GPCR	2.4×10^{-5}	0.9939
Molecules associated with elastic fibres	2.6×10^{-5}	0.0072929
DAP12 interactions	2.6×10^{-5}	0.78273
Cytochrome P ₄₅₀ - arranged by substrate type	3.2×10^{-5}	0.87019
GPCR ligand binding	3.8×10^{-5}	0.99417
Acylic chain remodelling of PC	4.0×10^{-5}	0.65415
Response to elevated platelet cytosolic Ca ²⁺	4.2×10^{-5}	0.55461
<i>Arachidonic acid metabolism</i>	4.4×10^{-5}	0.060298
Defective B4GALT7 causes EDS, progeroid type	4.9×10^{-5}	0.15497
Defective B3GAT3 causes JDSSDHD	4.9×10^{-5}	0.15497
Elastic fibre formation	4.9×10^{-5}	0.0019227
HS-GAG degradation	6.2×10^{-5}	0.017747
Bile acid and bile salt metabolism	6.2×10^{-5}	0.15497
Netrin-1 signalling	7.1×10^{-5}	0.95056
Integration of energy metabolism	7.1×10^{-5}	0.0019287
DAP12 signalling	7.9×10^{-5}	0.67835
GPCR downstream signalling	8.1×10^{-5}	0.88678
Diseases associated with glycosaminoglycan metabolism	8.7×10^{-5}	0.017747
Diseases of glycosylation	8.7×10^{-5}	0.017747
Signalling by Retinoic Acid	8.7×10^{-5}	0.13592
Signalling by Leptin	8.7×10^{-5}	0.15497
Signalling by SCF-KIT	8.7×10^{-5}	0.73399
Opioid Signalling	8.7×10^{-5}	0.99417
Signalling by NOTCH	0.0001	0.26453
Platelet homeostasis	0.0001	0.55912
Signalling by NOTCH1	0.00011	0.13797
Class B/2 (Secretin family receptors)	0.00011	0.4659
Diseases of Immune System	0.00013	0.15497
Diseases associated with the TLR signalling cascade	0.00013	0.15497
A tetrasaccharide linker sequence is required for GAG synthesis	0.00013	0.33566
Nuclear Receptor transcription pathway	0.00016	0.22735
Formation of Fibrin Clot (Clotting Cascade)	0.00016	0.0054639
Syndecan interactions	0.00016	0.3974
Class A/1 (Rhodopsin-like receptors)	0.00016	0.99454
HS-GAG biosynthesis	0.0002	0.37199
Platelet degranulation	0.0002	0.39003
EPH-ephrin mediated repulsion of cells	0.00021	0.6193

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

screens are costly, laborious, and specific to genetic background, computational analysis can augment candidate triage to either reduce the initial panel of screened genes or prioritise validation.

GPCR pathways were detected among both computational and experimental synthetic lethal candidates, with more support in the experimental screen (Table 4.8). The homogeneous cell line model may be more likely to detect particular pathways. For instance, SLIPT identified immune pathways, not expected to be detected in isolated cell culture. GPCR signalling was supported in experimental models Telford *et al.* (2015) with some of these pathways replicated in varied genetic backgrounds of patient samples. These pathways require further investigation such as identification of more specific pathways, higher order interactions, and modes of resistance.

The pathway composition across computational and experimental synthetic lethal candidates was informative with over-representation (Table 4.6) and supported by resampling analysis (Table 4.8), despite a modest intersection of genes between them (Figure 4.2). Either approach may be significant for a pathway in this intersection without being supported by the other: resampling analysis may support pathways that were not over-represented due to small effect sizes, thus both tests are required for a candidate pathway. The pathways detected by both over-representation and resampling are the strongest candidates for further investigation, such as $G_{\alpha s}$ signalling, a strong candidate in prior analyses with a role in the regulation of translation in cancer Gao and Roux (2015), another function supported by SLIPT analysis.

The predicted synthetic lethal partners occurred across functionally distinct pathways, including characterised functions of *CDH1*. This diversity is consistent with the wide ranging role of *CDH1* in cell-cell adhesion, cell signalling, and the cytoskeletal structure of epithelial tissues. Pathway structure may be relevant to identifying potential drug targets from gene expression signatures, indicating downstream effector genes and mechanisms leading to cell inviability. These distinct synthetic lethal gene clusters and pathways may further lead to the elucidation of drug resistance mechanisms.

4.3 Metagene Analysis

The gene signatures (Gatza *et al.*, 2011, 2014) were used to demonstrate the utility of the metagene approach for use on a wider range of pathways as was performed with the Reactome (Croft *et al.*, 2014) pathways as an alternative approach to identification of synthetic lethal pathways. Metagenes serve as a summary of activity for each pathway. The direction of metagenes (derived by the singular value matrix decomposition) is

generally arbitrary but care has been taken to ensure that these occur in a direction which reflect overall activation of the pathway (as described in Section 2.2.3). Metagenes were derived for well characterised gene signatures in breast cancer (Gatza *et al.*, 2011, 2014) to verify that that these pathway signatures are consistent with expected molecular properties of each molecular subtype (Parker *et al.*, 2009; Perou *et al.*, 2000). This was performed by examining the pathway expression of these breast cancer gene signatures in TCGA expression data. These metagenes were also compared to somatic mutation to evaluate mutation as a measure of gene activity in comparison to gene and protein expression.

The gene signatures (Gatza *et al.*, 2011, 2014) were used to demonstrate to utility of the metagene approach for use on a wider range of pathways. Having established that metagenes generated with this procedure reflect gene activity, the metagene procedure (in Section 2.2.3) was then applied to the Reactome pathways (Croft *et al.*, 2014). These Reactome metagenes were used for synthetic lethal analysis of pathways with SLIPT, directly using pathway activity for identifying synthetic lethal pathways with *CDH1*.

4.3.1 Pathway Expression

Pathway metagenes (generated as described in Section 2.2.3) for gene signatures of key processes in breast cancer (Gatza *et al.*, 2011) were used to check that metagenes were generated in the correct direction to indicate pathway activation. Some of these gene signatures are plotted in Figure 4.8 for comparison with clinical factors and somatic mutations. The “intrinsic subtype” was computed by performing the PAM50 procedure Parker *et al.* (2009) for RNASeq data which was highly concordant ($\chi^2 = 1305.9$, $p = 2.73 \times 10^{-268}$) with the subtypes provided by University of California, Santa Cruz (UCSC) for TCGA samples (TCGA, 2012) previously analysed by microarrays (as shown in Appendix D). Somatic mutations were reported for recurrently mutated genes in breast cancer, as reported by TCGA (TCGA, 2012), related genes, and those previously discussed to be important in hereditary breast cancers (*BRCA1*, *BRCA2*, and *CDH1*).

These gene signatures reflect intrinsic subtypes as expected. In particular, the oestrogen and progesterone receptor signatures are low in the predominantly ER⁻ and PR⁻ basal-like subtype tumours. These tumours also had the highest frequency of *TP53* mutations and a corresponding reduction of p53 metagene activity, as expected for loss of a tumour suppressor. The luminal A and luminal B tumour subtypes are the

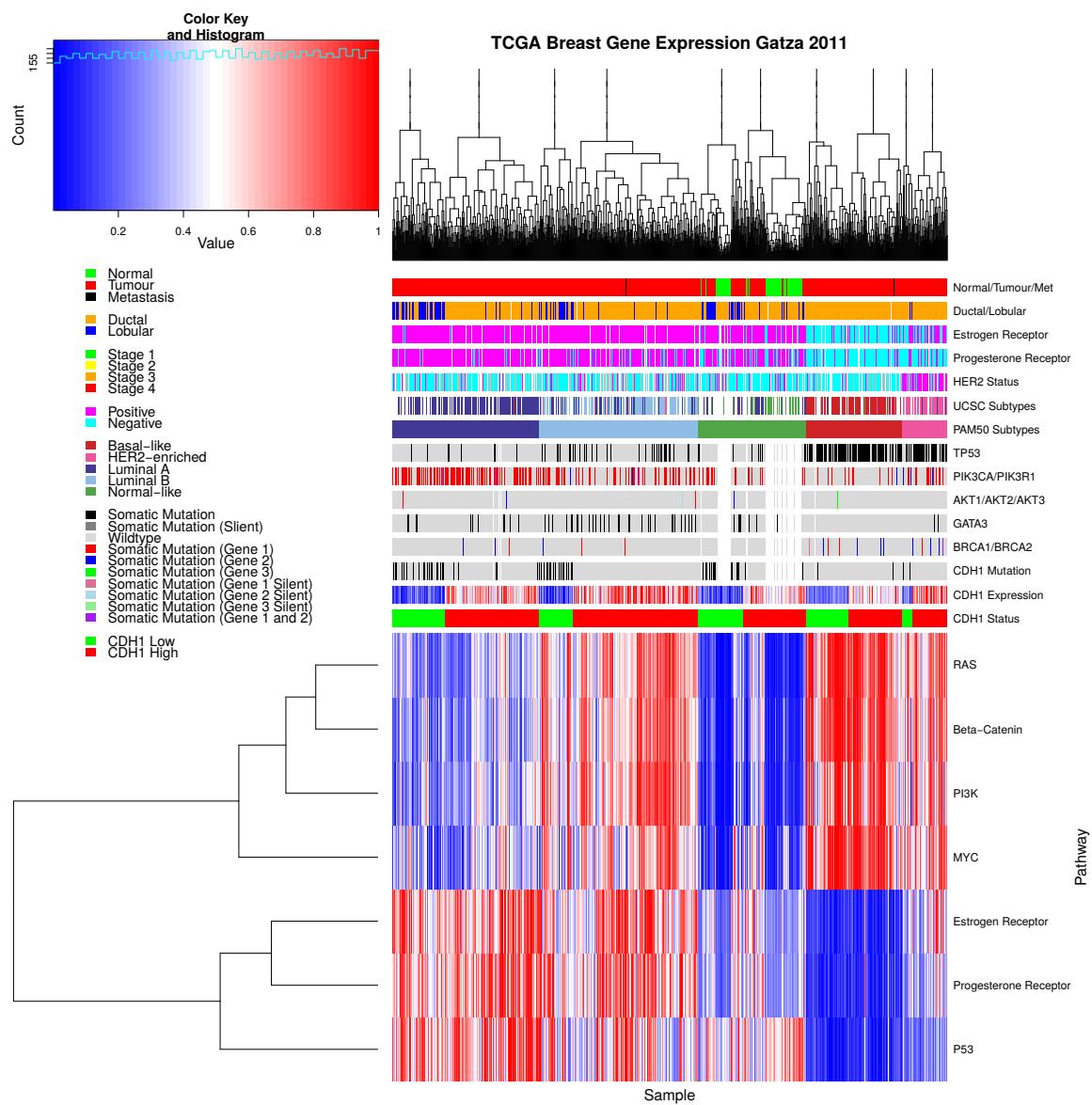


Figure 4.8: **Pathway metagene expression profiles.** Expression profiles for metagene signatures from Gatza *et al.* (2011) in TCGA breast data, annotated for clinical factors (with sample types and histological results coloured according to the legend) and cancer gene mutations (Negative values for mutation are light grey with missing data in white). Intrinsic subtypes are shown as derived from microarray (UCSC) and RNA-Seq (PAM50) data (Parker *et al.*, 2009; TCGA, 2012). Samples were clustered independently for each intrinsic subtype and by *CDH1* expression status. Pathway expression signatures are consistent with mutations and clinical subgroups.

most similar, which is reflected in these metagenes signatures, although they are distinguishable molecular subtypes as shown by elevated PI3K, AKT, RAS, and β -catenin signalling in luminal B tumours. However, these pathways were also elevated in basal-like and HER2-enriched subtypes and lowly expressed in the “normal-like” subtype (which contained the normal samples). These intrinsic subtype specific gene signature profiles were further supported with metagenes for an extended set of signatures (Gatza *et al.*, 2014), as shown in Figure C.9.

TP53 mutations were the most frequent and more common in the basal-like subtype. Similarly, *GATA3* mutations were more common in luminal subtype tumours. PI3K mutations were more frequent across breast tumours, although these were less common in the basal-like subtype despite an elevated metagene (this discrepancy will be discussed further in Section 4.3.2). *CDH1* mutations similarly occurred across molecular subtypes with the exception of the basal-like subtype (as observed in gene expression with Figure 4.1). *CDH1* low samples occurred in all subtypes but were predominantly of the lobular histological subtype. Apart from these genes, mutations did not show clear specificity to a particular subtype and the variation between samples reflects the range of molecular cascades that can result in tumours with similar molecular profiles, supporting the use of gene expression data for cancer diagnostics and identification of molecular targets.

The direction of each metagene was consistent with the clinical characteristics, which formed a consensus of gene activity as shown for the PI3K and ER signatures (Gatza *et al.*, 2011) in Figures 4.9 and 4.10, respectively. Supporting data for p53 and BRCA metagenes (Gatza *et al.*, 2011, 2014) are given in the Appendix (Figures C.10 and C.11). In each of the examples for gene signatures, the expression of the majority of the genes were highly concordant with the metagene, being either positively or negatively correlated. These were generally consistent with established clinical and molecular subtypes of breast cancer and the recurrent mutations shown. However, the *PIK3CA* and *PIK3R1* mutant samples did not necessarily have elevated PI3K pathway metagene activity (as shown in Figure 4.9).

4.3.2 Somatic Mutation

It should be noted that metagenes, while consistent with the consensus of constituent expressed genes, were not necessarily reflecting the somatic mutation status. The PI3K (Gatza *et al.*, 2011) metagene levels in particular, were not statistically significantly varying between mutant and wildtype *PIK3CA* samples (shown in Figure 4.11). How-

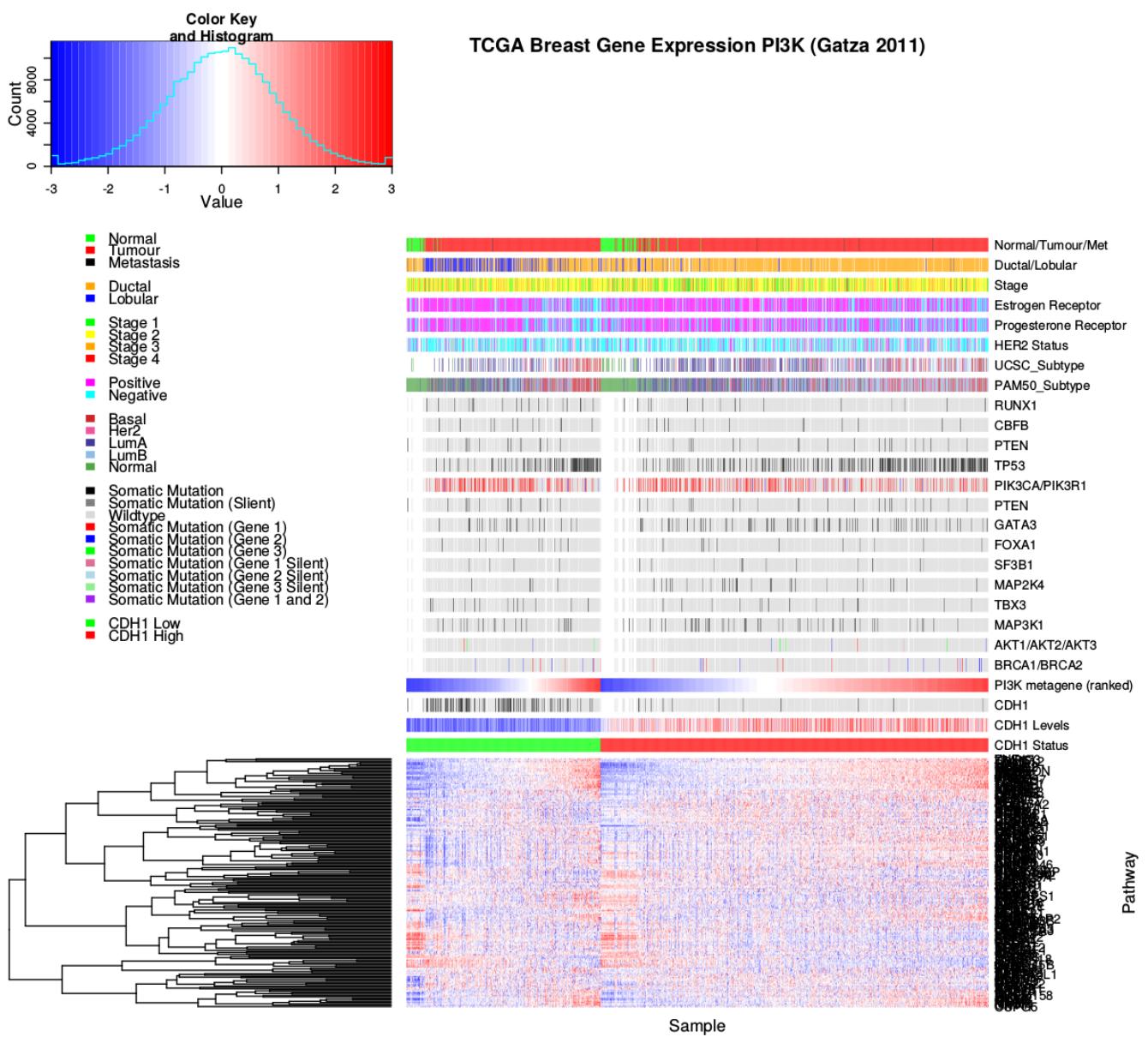
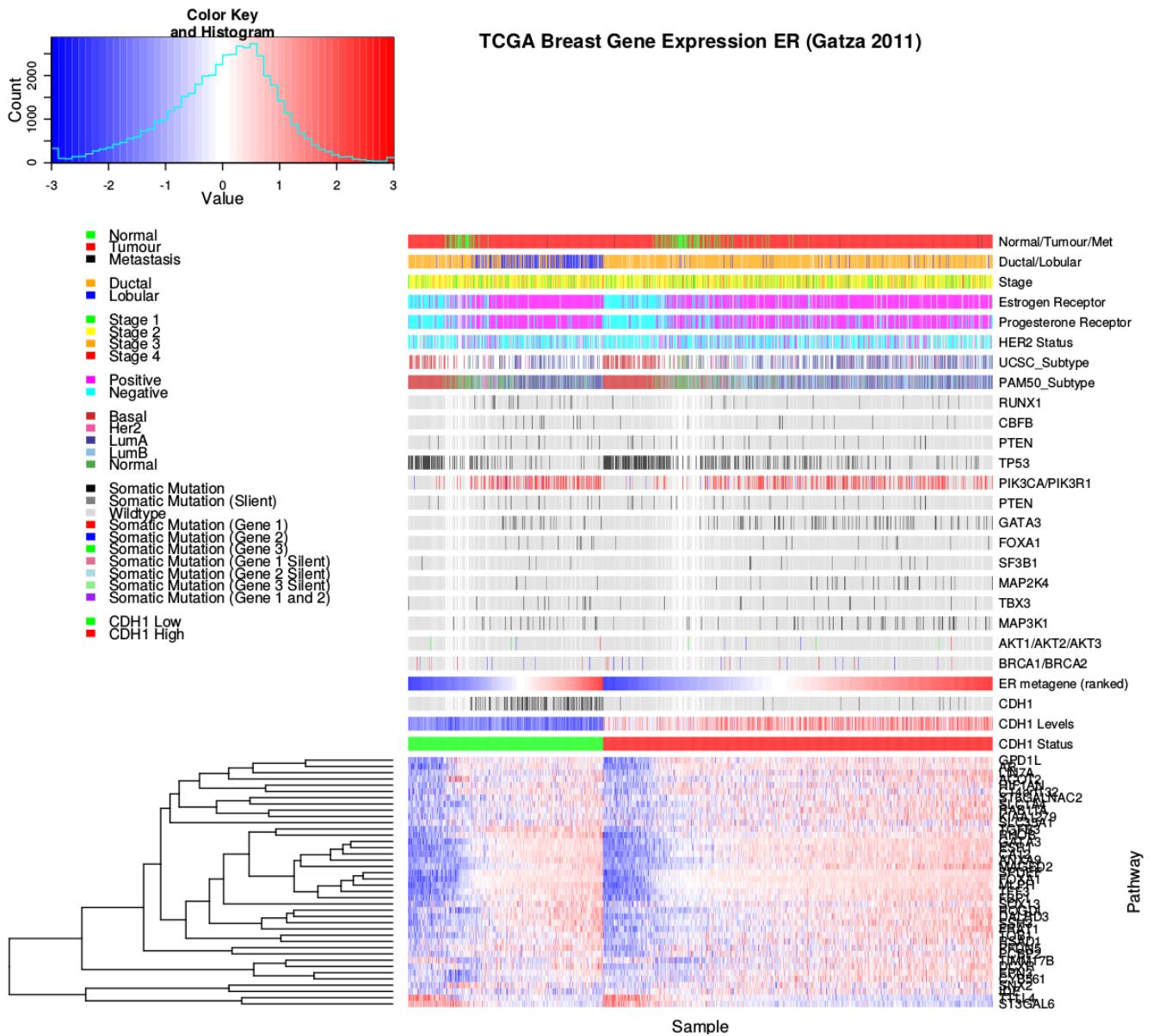


Figure 4.9: **Expression profiles for constituent genes of PI3K.** Expression profiles the genes contained in the PI3K gene signature from Gatza *et al.* (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by *CDH1* expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the PI3K metagene, although considerable proportion were inversely correlated with the metagene. Normal samples had low PI3K metagene expression and *TP53* mutant samples had high PI3K expression. Although, oncogenic *PIK3CA* and tumour suppressor *PIK3R1* mutations across samples including those with low metagene response.



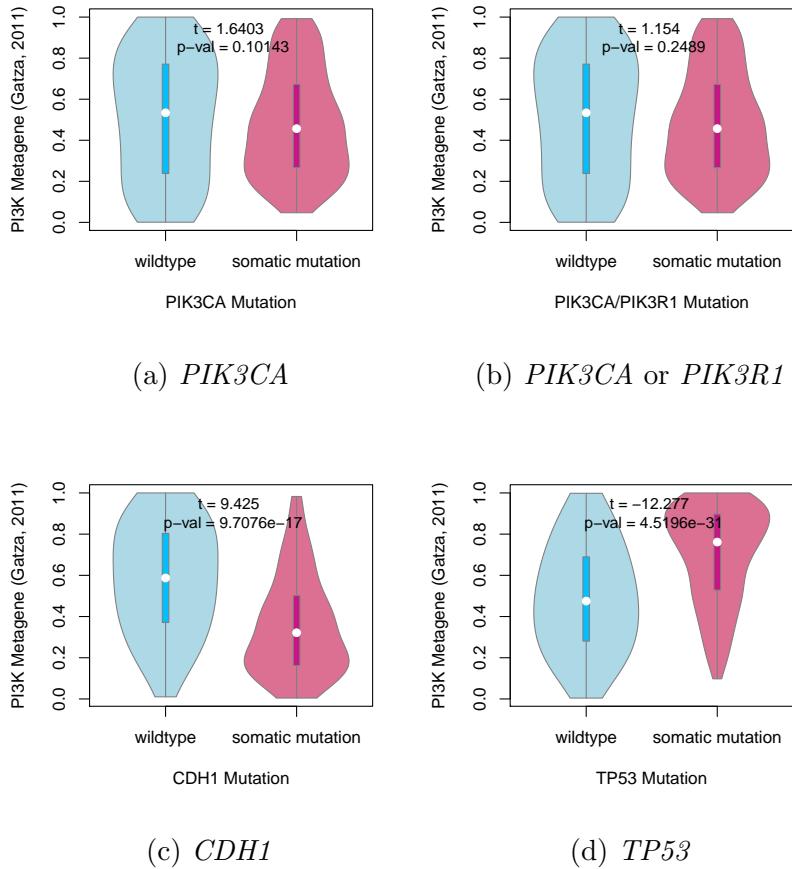


Figure 4.11: **Somatic mutation against the PI3K metagene.** Mutations in *PIK3CA*, *PIK3R1*, *CDH1*, and *TP53* were examined in TCGA breast cancer for their association with the PI3K (Gatza *et al.*, 2011) pathway metagene. The tumour suppressors *CDH1* and *TP53* showed an increase and decrease in the metagene respectively, whereas *PIK3CA* and *PIK3R1* mutations had little effect on the metagene levels.

ever, the PI3K metagene differed across *CDH1* and *TP53* mutations, remarkably in opposite directions considering that PI3K is an oncogenic growth pathway and these are both most frequently tumour suppressors inactivated in cancers. This shows that *CDH1* and *TP53* deficient tumours have distinct molecular growth pathways and that synthetic lethal interventions against loss of *CDH1* function may not be applicable to other cancers with driver mutations such as *TP53*, although these were kept in the analysis for comparison. These differences may be related to these mutations being more frequent in tumours with difference clinical characteristics (as observed in Section 4.3.1). Thus mutations do not necessarily have corresponding changes in pathway

expression, particularly for oncogenes which may change in function rather than being upregulated.

While the more specific *PIK3CA* (Gatza *et al.*, 2014) metagene showed significant differences with *PIK3CA* and *PIK3R1* mutations (as shown in Figure C.6), this metagene replicated stronger differences for *CDH1* and *TP53*. These differences were less pronounced in the protein levels of p110 α (enocded by *PIK3CA*) and the downstream AKT gene (shown in Figures C.7 and C.8 respectively). However, this may be due to this regulatory cascade (kinases) being transmitted as a change in protein state (phosphorylation) rather than changes in expression levels. Another consideration is that mutations at different loci have different effects on protein function, particularly for oncogenes.

4.3.3 Synthetic Lethal Pathway Metagenes

Pathway metagenes for Reactome pathways (generated as described in Section 2.2.3) were also used for testing synthetic lethal partner pathways with *CDH1* by SLIPT. Since the metagenes have are higher when the pathway as a whole is activated, they are amenable to SLIPT analysis using low metagene levels for inactivated pathways. These synthetic lethal metagenes differed to the over-represented pathways among synthetic lethal gene candidates. However, there were some similarities to previous findings, as shown in Tables 4.9. In particular, translational pathways were replicated as observed in Table 4.2. While the specific pathways differ, immune pathways (e.g., NF- κ B) were also supported by metagene synthetic lethal analysis.

Signalling pathways were more strongly supported by mtSLIPT analysis of metagene pathway expression against *CDH1* mutation, as shown in Table C.7, although these results were generally less statistically significant than expression analyses. Signalling pathways detected as synthetic lethal metagenes include G $_{\alpha z}$, insulin-related growth factor (IGF), GABA receptor, G $_{\alpha s}$, S6K1 and various toxin responses mediated by GPCRs. Metabolic processes including processing of carbohydrates and fatty acids were also implicated across these analyses.

The metagene analyses differ more between expresssion and *CDH1* mutation than previous analyses, with more specific signalling pathways identified in the mutation analysis. This supports the usage of a complete null mutant model in experimental testing for synthetic lethality of signalling pathways against *CDH1* inactivation rather than a knockdown in expression. However, low expression of partners has been used in either case to be applicable to dose-dependent pharmacological inhibition and across

Table 4.9: Candidate synthetic lethal metagenes against *CDH1* from SLIPT

Pathway	ID	Observed	Expected	χ^2 value	p-value	p-value (FDR)
Glycogen storage diseases	3229121	68	130	176	6.62×10^{-37}	1.82×10^{-34}
Myoclonic epilepsy of Lafora	3785653	68	130	176	6.62×10^{-37}	1.82×10^{-34}
Diseases of carbohydrate metabolism	5663084	68	130	176	6.62×10^{-37}	1.82×10^{-34}
Arachidonic acid metabolism	2142753	81	130	157	8.13×10^{-33}	1.49×10^{-30}
Translation initiation complex formation	72649	70	130	152	7.08×10^{-32}	1.17×10^{-29}
Synthesis of 5-eicosatetraenoic acids	2142688	68	130	151	1.25×10^{-31}	1.88×10^{-29}
SRP-dependent cotranslational protein targeting to membrane	1799339	69	130	150	2.01×10^{-31}	2.76×10^{-29}
L13a-mediated translational silencing of Ceruloplasmin expression	156827	72	130	148	5.91×10^{-31}	6.44×10^{-29}
3' -UTR-mediated translational regulation	157279	72	130	148	5.91×10^{-31}	6.44×10^{-29}
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	72662	70	130	147	1.14×10^{-30}	9.28×10^{-29}
Formation of the ternary complex, and subsequently, the 43S complex	72695	70	130	147	1.14×10^{-30}	9.28×10^{-29}
Ribosomal scanning and start codon recognition	72702	70	130	147	1.14×10^{-30}	9.28×10^{-29}
Eukaryotic Translation Elongation	156842	72	130	146	1.19×10^{-30}	9.28×10^{-29}
Nonsense Mediated Decay independent of the Exon Junction Complex	975956	71	130	146	1.24×10^{-30}	9.28×10^{-29}
Viral mRNA Translation	192823	70	130	146	1.51×10^{-30}	1.04×10^{-28}
Eukaryotic Translation Termination	72764	70	130	146	1.51×10^{-30}	1.04×10^{-28}
NF- κ B is activated and signals survival	209560	71	130	145	1.90×10^{-30}	1.19×10^{-28}
Peptide chain elongation	156902	72	130	145	1.91×10^{-30}	1.19×10^{-28}
Influenza Life Cycle	168255	70	130	145	1.95×10^{-30}	1.19×10^{-28}
Formation of a pool of free 40S subunits	72689	73	130	145	2.01×10^{-30}	1.19×10^{-28}
Nonsense-Mediated Decay	927802	71	130	145	2.44×10^{-30}	1.34×10^{-28}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	975957	71	130	145	2.44×10^{-30}	1.34×10^{-28}
GTP hydrolysis and joining of the 60S ribosomal subunit	72706	72	130	145	2.58×10^{-30}	1.37×10^{-28}
Influenza Viral RNA Transcription and Replication	168273	72	130	144	4.01×10^{-30}	2.07×10^{-28}
Signalling by NOTCH1 HD Domain Mutants in Cancer	2691230	79	130	143	5.99×10^{-30}	2.82×10^{-28}

Strongest candidate SL partners for *CDH1* by SLIPT with observed and expected numbers of TCGA breast cancer samples with low expression of both *CDH1* and the metagene.

genes where mutations have different functional consequences, including variants of unknown significance.

These results show an independent pathway-based approach to detecting synthetic lethal gene functions interacting with *CDH1*. The use of synthetic lethal metagenes replicates support for these pathways independent of pathway size (as genes are weighted equally). Along with the verifying that the direction of metagenes recapitulates the activity of a pathway, these demonstrate that many of the pathways previously identified from over-represented synthetic lethal genes (detected by SLIPT) are synthetic lethal pathways with their activity dependent on synthetic lethal genes rather than containing synthetic lethal genes as inhibitors or peripheral regulators of the pathways.

4.3.4 Synthetic Lethality in Breast Cancer

The synthetic lethal analysis against low *CDH1* expression supports prior findings in translational and immune pathways even if they were not able to be detected in an experimental screen (Telford *et al.*, 2015). Together these findings support the role of *CDH1* loss in cancer disrupting cell signalling with wider effects on protein translation and metabolism necessary for the proliferation of cancer cells. This is consistent with

the GPCR pathways such as $G_{\alpha s}$ signalling being supported by SLIPT gene candidates and the experimental primary siRNA screen, as shown by resampling in Section 4.2.5.1.

4.4 Replication in Stomach Cancer

CDH1 is also important in stomach cancer biology as a driver tumour suppressor gene, including as a germline mutation in many cases of hereditary diffuse gastric cancer. The synthetic lethal analysis of genes and pathways (previously identified for TCGA breast cancer data) was replicated in TCGA stomach cancer. The accompanying data for SLIPT analysis against *CDH1* expression is provided in Appendix E.

While the sample size was lower for TCGA stomach cancer (particularly for mutations), these results serve to support the findings in breast cancer in an independent patient cohort and tissue samples. The molecular profiling, including RNA-Seq expression, were performed by TCGA using the sample procedures as for breast cancer and the findings reported here were performed used data analysis techniques identical to those presented previously. These procedures should ensure as close comparison as feasible across cancer types for those relevant to HDGC and recurrent *CDH1* mutations.

The strongest SLIPT genes for stomach cancer (shown in Table E.1) did not necessarily directly correspond to those observed in breast cancer (shown in Table 4.1). However, several gene functions were replicated in stomach cancer. Together, these gene candidates indicate widespread functions of *CDH1* and strongly detectable synthetic lethality with many genes from a strategy that can be applied across cancer types. More specifically, the signalling genes included GPCR signalling genes, which was one of the most supported synthetic lethal pathways in breast cancer analysis, the experimental screen (Telford *et al.*, 2015). These findings were further supported by the pathways over-represented in SLIPT candidates from TCGA stomach cancer (shown in Table E.2) which replicated the translational and immune pathways observed in TCGA breast cancer (shown in Table 4.2) and further supported GPCR signalling pathways, including the class A/1 receptors. The extracellular matrix was also detected at the pathway level in stomach cancer, including elastic fibres, glycosylation, collagen, and integrin cell-surface interactions. While fewer pathways were supported by resampling for the intersection of SLIPT and experimental screen (Telford *et al.*, 2015) candidate partners in stomach cancer than breast cancer, many of those detected (shown in Table E.6) replicate those detected in breast cancer (shown in Table 4.8). The pathways detected by both permutation and over-representation were more likely to be replicated across stomach and breast cancer than those detected by over-representation

alone, supporting the use of this procedure to detect synthetic lethal pathways applicable across cancer types. These include G_{αs} signalling and elastic fibre formation as discussed for breast cancer (in Section 4.2.5.1).

4.5 Discussion

4.5.1 Strengths of the SLIPT Methodology

Synthetic lethal discovery with SLIPT used established statistical procedures to identify putative partner genes from gene expression data. Such use of the χ^2 -value is amenable to pathway or permutation analyses and could feasibly be applied to other disease gene or pair-wise across the genome, although genome-wide approaches were unable to find informative candidate genes for E-cadherin (Lu *et al.*, 2015). Synthetic lethal discovery in cancer has focused on genes with severe cellular mutant phenotypes, such as essential genes or the oncogenes *TP53* and *AKT* (Lu *et al.*, 2015; Tiong *et al.*, 2014; Wang and Simon, 2013), with other cancer genes, such as *CDH1*, requiring more focused investigations. Prior computational approaches for synthetic lethal discovery, in cancer, vary widely (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015; Tiong *et al.*, 2014; Wappett *et al.*, 2016). There is no consensus as to which approach is more appropriate, and the methods are difficult to compare, as they either do not have a released code implementation or do not make predictions solely from normalised expression data.

However, the query-based approach demonstrated by SLIPT analysis is suitable for wider application on expression data and for augmenting experimental studies such as high-throughput screens. This approach has identified biologically plausible synthetic lethal pathways for *CDH1*, triaged candidates from experimental screening (Telford *et al.*, 2015), and replicates genes and pathways across breast and stomach cancer datasets. In addition, SLIPT avoids critical assumptions underlying the design of some approaches such as co-expression of synthetic candidates or that interacting gene pairs will have known (annotated) similarities in function.

The DAISY methodology Jerby-Arnon *et al.* (2014), which took a similar query-based approach with the tumour suppressor *VHL*, has been critiqued for being too stringent (Lu *et al.*, 2015) which impedes pathway analysis. Since functional redundancy does not require genes to be expressed at the same time, the SLIPT approach does not assume co-expression of synthetic lethal genes which may enrich for synthetic lethal genes in established coregulated pathways. Rather, the interpretation of synthetic lethality for SLIPT was similar to other computational methods based on

‘co-loss under-representation’, ‘compensation’, or ‘simultaneous differential expression’ (Lu *et al.*, 2015; Tiong *et al.*, 2014; Wang and Simon, 2013).

Genomics analyses are prone to false-positives and require statistical caution, particularly where working with gene-pairs scale up the number of multiple tests drastically, at the expense of statistical power. Experimental screens for synthetic lethality are also error-prone (Feces de la Cruz *et al.*, 2015; Lord *et al.*, 2015; Lu *et al.*, 2015), especially with false-positives, raising the need for understanding the expected behaviour and number of functional relationships and genetic interactions in the genome, or in discovery of synthetic lethal partners of a particular query gene. Thus analyses throughout this thesis have focused on querying for partners of a particular gene of interest. Statistical modelling and simulations (in Section 3.3 and Chapter 6) will further support the design decisions underlying SLIPT analysis and its strengths over other approaches.

4.5.2 Synthetic Lethal Pathways for E-cadherin

Specific genes were difficult to replicate across experiments. This is consistent with gene expression profiles for synthetic lethal partners reflecting the complexity of biological pathways which are subject to higher-order interactions and do not consistently compensate for loss of gene function across all samples (Jerby-Arnon *et al.*, 2014; Kelly, 2013; Lu *et al.*, 2015). The predicted synthetic lethal partners of *CDH1* (with FDR correction) were investigated with gene expression profiles and clinical variables to find relationships in gene expression, gene function, and clinical characteristics. The large number of genes detected indicates that synthetic lethal detection is potentially error-prone, and that identifying genes relevant for clinical application will be difficult without a supporting biological pathway rationale. As such, investigations into the genes identified by SLIPT, the correlation structure between them, and those which were validated by experimental screening (Telford *et al.*, 2015) focused at the pathway level throughout this Chapter. Similarly, comparisons across analyses were largely made at the pathway level, including comparisons between expression and mutation, breast and stomach TCGA datasets.

Potential synthetic lethal partners of *CDH1* identified by SLIPT had many distinct functions, with each gene cluster highly expressed in different patient subgroups (Figure 4.1). The expression profiles of the SL partners of *CDH1* predicted from TCGA breast cancer RNA-Seq data (expected to have compensating high or stable expression) and their corresponding functional enrichment found in subgroups of genes, particularly

among *CDH1* low breast tumours. Ductal breast cancers showed higher expression of synthetic lethal partners suggesting treatment would be more effective in this tumour subtype. However, there was consistently low expression of SL partners in estrogen receptor negative tumours, although this is independent of tumour stage and consistent with poor prognosis in these patients and could inform other treatment strategies or prevent ineffective treatment further impacting quality of life in these patients. These results suggest that synthetic lethal partner expression varies between patients; that these different tumour classes would react differently to the same treatment; that treatment of different pathways and combinations in different patients is the most effective approach to target genes compensating for *CDH1* gene loss; and that the expression of synthetic partners could be a clinically important biomarker.

The pathways that synthetic lethal partners of *CDH1* identified by SLIPT were involved in a diverse range of biological functions and differed to those detected experimentally. This discrepancy may be accounted for by gene expression analyses detecting both synthetic lethal partners, as screened for experimentally by Telford *et al.* (2015), and their downstream targets (not detected by siRNA), capturing the wider pathways and mechanisms involved in synthetic lethality with *CDH1* inactivation. In particular, GPCR phosphorylation cascades (which regulate gene expression and translation in cancers (Gao and Roux, 2015)) were predicted to be synthetic lethal with *CDH1*. The predicted synthetic lethal partners occurred across functionally distinct pathways, including characterised functions of *CDH1*. The most consistently supported pathways included elastic fibres in the extracellular matrix, GPCR signalling, and translation presenting vulnerabilities for *CDH1* deficient cancer cells from extracellular stimuli to the core growth mechanisms of a cell.

This diversity in synthetic lethal functions is consistent with the wide ranging role of *CDH1* in cell-cell adhesion, cell signalling, and the cytoskeletal structure of epithelial tissues. Pathway structure may be relevant to identifying potential drug targets from gene expression signatures, indicating downstream effector genes and mechanisms leading to cell inviability. Identification of distinct synthetic lethal gene clusters may further lead to the elucidation of drug resistance mechanisms. While these pathways are indicative of the main functions of E-cadherin and synthetic lethal partners, it remains to identify the genes within these pathways that are the most actionable or supported across SLIPT analysis in patient samples and detected by experiments in preclinical models (Chen *et al.*, 2014; Telford *et al.*, 2015). The specific genes within key pathways will be discussed in Chapter 5, along with further investigations into

their relation to pathway structure. While these are important clinical implications, the synthetic lethal predictions lack enough confidence for direct translation into pre-clinical models or clinical applications leading to a need for statistical modelling and simulation of synthetic lethality in genomics expression data.

These synthetic lethal pathways have potential clinical implications, particularly those supported in pre-clinical models and in patient expression data. However, further validation of gene candidates will be necessary to ensure that these are able to be reproduced in further pre-clinical studies, they are applicable to tumours *in vivo*, and that effective inhibitory agents can be repurposed or designed against them.

4.5.3 Replication and Validation

4.5.3.1 Integration with siRNA Screening

The pathway composition across computational and experimental synthetic lethal candidates was informative with over-representation (Table 4.6) and supported by resampling analysis (Table 4.8), despite a modest intersection of genes between them (Figure 4.2). Either approach may be significant for a pathway in this intersection without being supported by the other: resampling analysis may support pathways that were not over-represented due to small effect sizes, thus both tests are required for a candidate pathway.

The pathways detected by both over-representation and resampling are the strongest candidates for further investigation and the pathway structure analyses in Chapter 5 will focus on these pathways detected by both over-representation and resampling. Particularly, those replicated across datasets or with pathway metagenes. In addition to GPCR pathways detected across these analyses, the PI3K cascade will also be investigated in Chapter 5, this signalling pathway is a well characterised mediator between GPCR receptors and regulation of translation (Gao and Roux, 2015) (both detected throughout this Chapter) and exhibited unexpected behaviour with pathway metagenes (in Section 4.3). This pathway is activated by protein Phosphorylation states and thus inactivation may not be detectable with expression.

However, the SLIPT approach was shown to be predictive of which siRNA primary screen candidate partners of *CDH1* were validated in a secondary screen (as shown in Section 4.2.4). These results further support SLIPT for identifying robust synthetic lethal candidates which can be validated and as a triage approach for interpreting screening experiments.

4.5.3.2 Replication across Tissues

Furthermore, synthetic lethal partners identified by SLIPT were replicated across breast and stomach cancer. These were particularly concordant at the pathway level, as expected between tissues since synthetic lethal pathways have higher conservation between species (Dixon *et al.*, 2008). These findings support gene functions conserved across *CDH1* deficient cancers in breast and stomach tissues, presenting vulnerabilities that could be applied against molecular targets in both cancers. In addition, these analyses serve as a replication across independent patient cohorts from breast and stomach cancers, decreasing the likelihood of the synthetic lethal pathways detected being false positives or artifacts of either dataset.

Synthetic lethal pathways were also replicated across expression analyses of TCGA patient samples in heterogeneous tumours and homogeneous cell line isolates. This further supports that the subset of synthetic lethal functions detectable in experimental models (Chen *et al.*, 2014; Telford *et al.*, 2015) would be applicable tumours of patients with *CDH1* deficient cancers.

There are many gene functions replicated across breast cancer gene expression analyses. Many of these were also replicated with mutation analysis and with stomach cancer or cell line expression data. These pathways were more consistent across replication analyses than previous investigations with TCGA microarray data (Kelly, 2013).

4.6 Summary

We have developed a simple, interpretable, computational approach to predict synthetic lethal partners from genomics data. The analyses focus on gene expression data as it is widely available for applications in other cancers and other disease genes, particularly those with malignant loss of function.

This approach has been applied to robustly detect synthetic lethal pathways for the E-cadherin (*CDH1*) in TCGA breast cancer molecular profiles with comparisons to experimental screening (Telford *et al.*, 2015) in cell lines, and replication in TCGA stomach cancer molecular profiles and across cell types in the cancer cell line encyclopaedia. The pathway replicated across several analyses included extracellular matrix pathways (e.g., elastic fibres formation), cell signalling (including GPCRs), and core gene regulation and translation processes crucial for the growth and proliferation of cancer cells. These pathways show evidence of non-oncogene addiction for *CDH1* deficient cells and present vulnerabilities which may be exploited for specific treatment against *CDH1* mutations in HCGC and sporadic cancers. There was also support for

synthetic lethal pathways with *CDH1* in cell adhesion and cytoskeletal processes to which *CDH1* belongs, supporting the finding that synthetic lethality occurs within biological pathways (Boone *et al.*, 2007; Kelley and Ideker, 2005).

While translational and immune pathways detected by SLIPT were not supported by primary siRNA screening (Telford *et al.*, 2015), these were replicated across various analyses. Due to the differences between an experimental cell line model (Chen *et al.*, 2014; Fece de la Cruz *et al.*, 2015) and patient molecular profiles (Bass *et al.*, 2014; TCGA, 2012), these would not be expected to be completely concordant. Furthermore, many pathways are difficult to test in an isolated experimental system. Nevertheless, many of the genes and pathways detected by SLIPT are suitable to inform further investigations and triage of therapeutic targets against *CDH1* deficient tumours in combination with experimental screening.

A characteristic of gene interaction networks is a scale-free topology leading to highly interacting hub genes, these represent important genes in a functional network. Cell surface interactions, the extracellular matrix, and cell signalling (particularly PI3K/AKT signalling) were also found to be synthetic lethal hubs with more interactions detected than other genes. This indicates that these pathways are functionally important to survival of cancer cells since they are subject to high functional redundancy, despite frequent disruptions in cancer. These pathways being involved in a disproportionate number of synthetic lethal interactions is also consistent with their detection for *CDH1*.

Thus synthetic lethal pathways have been identified using TCGA patient molecular profiles and experimental screening results. Some of these were robustly replicated across these datasets and against *CDH1* mutation or expression analysis. However, there remains the need to identify actionable genes within these pathways, relationships with experimental candidates, and how these pathways may affect viability when lost. While the genes identified between these analyses were less concordant the results of the TCGA breast cancer analysis will be used to test pathway structure relationships and further examine the synthetic lethal genes detected in the following Chapter.

Chapter 5

Synthetic Lethal Pathway Structure

Having identified key pathways implicated in synthetic lethal genetic interactions with *CDH1* (in Chapter 4), these were investigated for the synthetic lethal genes within them and their relationships to pathway structure in Reactome pathways. This chapter will focus on the pathway structure of biological pathways detected across analyses in Chapter 4.

The synthetic lethal genes identified were further examined within the context of biological pathways. Specifically, investigations were performed on whether synthetic lethal candidates, detected by SLIPT or siRNA, exhibited differences with respect to network metrics of pathway structure of connectivity and importance in the network (as described in Sections 2.4.4 and 3.5.3). The relationships between synthetic lethal candidates, detected by either approach, were also considered to detect whether genes detected by SLIPT were upstream or downstream of genes detected by siRNA. These directional relationships were tested by resampling (as described in Sections 3.4.1 and 3.4.1.1) and comparisons to the pathway hierarchical score based on biological context (as derived in Section 3.4.1.2). Together these investigations into structural relationships demonstrate how a combination of network biology and statistical techniques can be performed with genes identified by a bioinformatics analysis.

5.1 Synthetic Lethal Genes in Reactome Pathways

The graph structure for Reactome pathways was obtained from Pathway Commons via BioPAX (as described in Section 2.4.2). The pathways describe the (directional) relationships between biomolecules, including genes that encode proteins in biological pathways. These relationships include cell signalling (e.g., kinase phosphorylation cascades), gene regulation (e.g., transcription factors, chromatin modifiers, RNA bind-

ing proteins), and metabolism (e.g., the product of an enzyme being the substrate of another). Together these relationships describe the known functional pathways in a human cell with a reasonable resolution, from a curated database supported by publications documenting pathway relationships.

Pathway structures from the Reactome network (as described in Section 2.4.3) were used to derive the graph structure of each biological pathway. The synthetic lethal candidate genes for notable pathways discussed in Chapter 4, including candidate synthetic lethal pathways of *CDH1*, were examined to show the SLIPT and siRNA candidates within these pathways. The synthetic lethal genes considered here are those candidates detected by SLIPT (as described in Section 3.1) in TCGA breast cancer expression and mutation data (TCGA, 2012) in comparison to the candidate gene partners from the siRNA screening in breast cell lines (Telford *et al.*, 2015).

5.1.1 The PI3K/AKT Pathway

The phosphoinositide 3-kinase (PI3K) cascade signalling pathway exhibited unexpected results with metagene analyses (as discussed in Section 4.3). This pathway is also of interest because mediating signals between the G protein coupled receptors and regulation of protein translation have both been strongly implicated to be synthetic lethal pathways with loss of *CDH1* function (in Chapter 4). These pathways have all subject to dysregulation in cancer (Courtney *et al.*, 2010; Dorsam and Gutkind, 2007; Gao and Roux, 2015). Thus the PI3K cascade will be examined along with the most supported synthetic lethal pathways (as identified in Chapter 4).

The PI3K pathway is also an ideal pathway in which to test pathway structure because it has an established direction of signal transduction from extracellular stimuli (and membrane bound receptors) to the inner mechanisms of the cell, namely, the regulation of protein translation. The production of proteins is necessary for the growth of the cell so it is reasonable to suggest that these processes may be subject to (non-oncogene) addiction in some cancer cells which rely upon them for sustained protein production and cell growth. This is also supported by the oncogenes *PIK3CA* and *AKT1* being involved with the PI3K cascade and related PI3K/AKT pathway which may be subject to oncogene addiction when these proto-oncogenes are activated.

The PI3K cascade was not supported across SLIPT in TCGA breast expression data and the siRNA primary screen by over-representation (in Section 4.2.5) or resampling (in Section 4.2.5.1) but genes were detectable by either approach (as shown in Figure 5.1). While few genes were identified by both approaches, these include genes

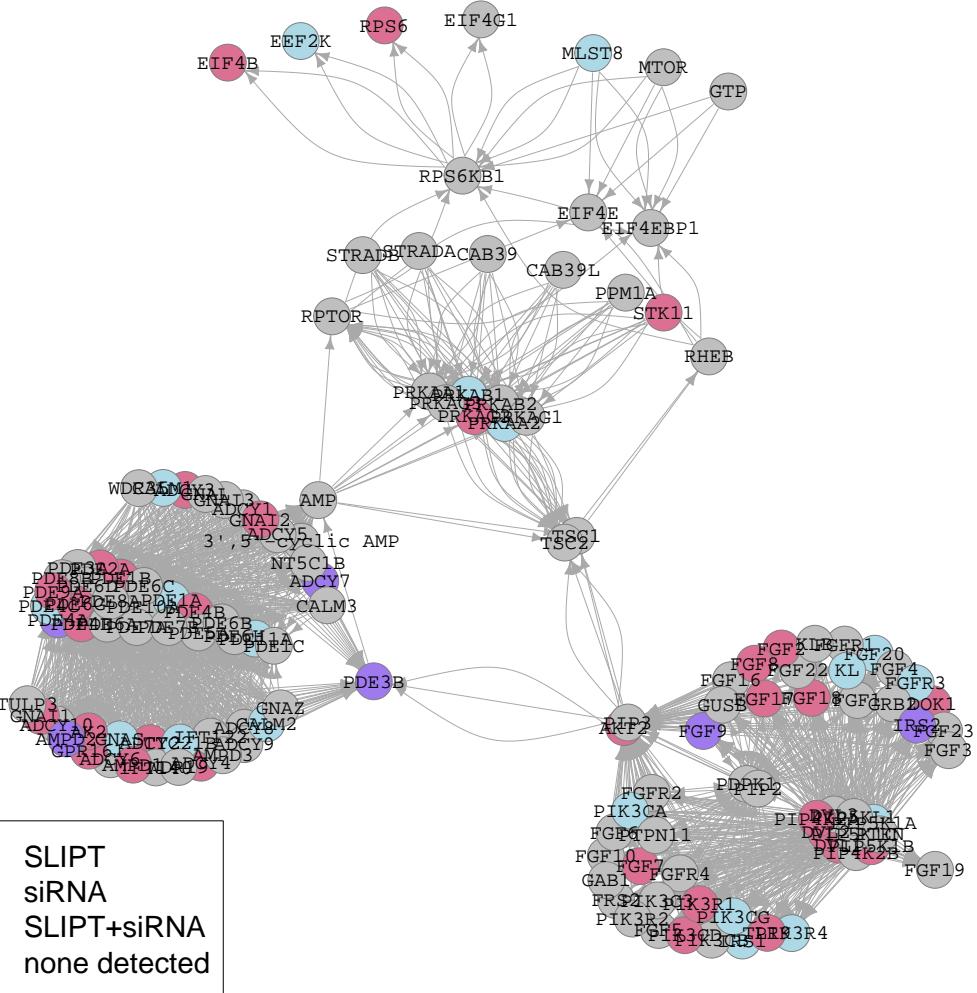


Figure 5.1: Synthetic Lethality in the PI3K Cascade. The Reactome PI3K Cascade pathway with synthetic lethal candidates coloured as shown in the legend.

that are highly connected in the PI3K cascade and are hubs to information transmission such as *FGF9*, *PDE3B*, and *PDE4A*. The key upstream genes *PIK3CA* and *PIK3CG* were detected by siRNA whereas the downstream *PIK3R1* and *AKT2* genes were detected by SLIPT. Gene detected by either method were also prevalent in the PI3K, phosphodiesterase (PDE), and AMP-activated protein kinase (AMPK) modules, in addition to the downstream translation factors and ribosomal genes (*EIF4B*, *EEF2K*, and *RPS6*). Together these suggest that there may be further structure between the SLIPT and siRNA candidate partners of *CDH1* in pathways as illustrated by PI3K. As

such, pathway structure will be investigated to detect differences in the upstream and downstream gene candidates of those detected by either method. Pathway structure may account for the disparity between SLIPT and siRNA genes, even in pathways such as PI3K where they did not significantly intersect. For instance, SLIPT gene partners may be downstream of siRNA candidates rather than replicating them directly.

This disparity between SLIPT and siRNA gene candidate synthetic lethal partners of CDH1, that is a high number of genes detected by either approach with few detected by both, was replicated in the related PI3K/AKT pathway and the “PI3K/AKT in cancer” pathway (shown in Appendix Figures F.1 and F.2). Many synthetic lethal candidates were at the upstream core of these pathway networks and the downstream extremities. It is particularly notable that the many genes important in cell signalling and gene regulation were detected by either synthetic lethal detection approach. These include *AKT1*, *AKT2*, and *AKT3*, the Calmodulin signalling genes *CALM1* and *CAMK4*, and the forkhead family transcription factors *FOXO1* (a tumour suppressor) and *FOXO4* (an inhibitor of EMT).

5.1.2 The Extracellular Matrix

The extracellular pathways “elastic fibre formation” and “fibrin clot formation” (shown in Figures 5.2 and 5.3 respectively) were both supported across analyses (in Chapter 4). Significant over-representation and resampling the intersection between SLIPT (for TCGA breast cancer) and siRNA gene candidates showed that both approaches identified these pathways.

Particularly for elastic fibres (Figure 5.2), the vast majority of genes were detected by either approach in addition to a significant proportion of genes detected by both approaches (as determined in Section 4.2.5). The genes detected by both approaches also appeared to have a non-random distribution in the network with *TFGB1*, *ITGB8*, and *MFAP2* exhibiting high connectivity, and having a central role in their respective pathway modules. In addition to a structural role in the extracellular matrix and connective tissue (including the tumour microenvironment), these proteins including Furin, transforming growth factor β (TGF β), and the bone morphogenic proteins (BMPs), are also involved in responses to endocrine signals and interact with the cellular receptors for signalling pathways. Therefore it is plausible that *CDH1* deficient tumours will be subject to non-oncogene addiction to the extracellular environment and growth signals arising from this pathway. The pathway structure also indicative for further investigation that the genes detected by siRNA (or both approaches) may

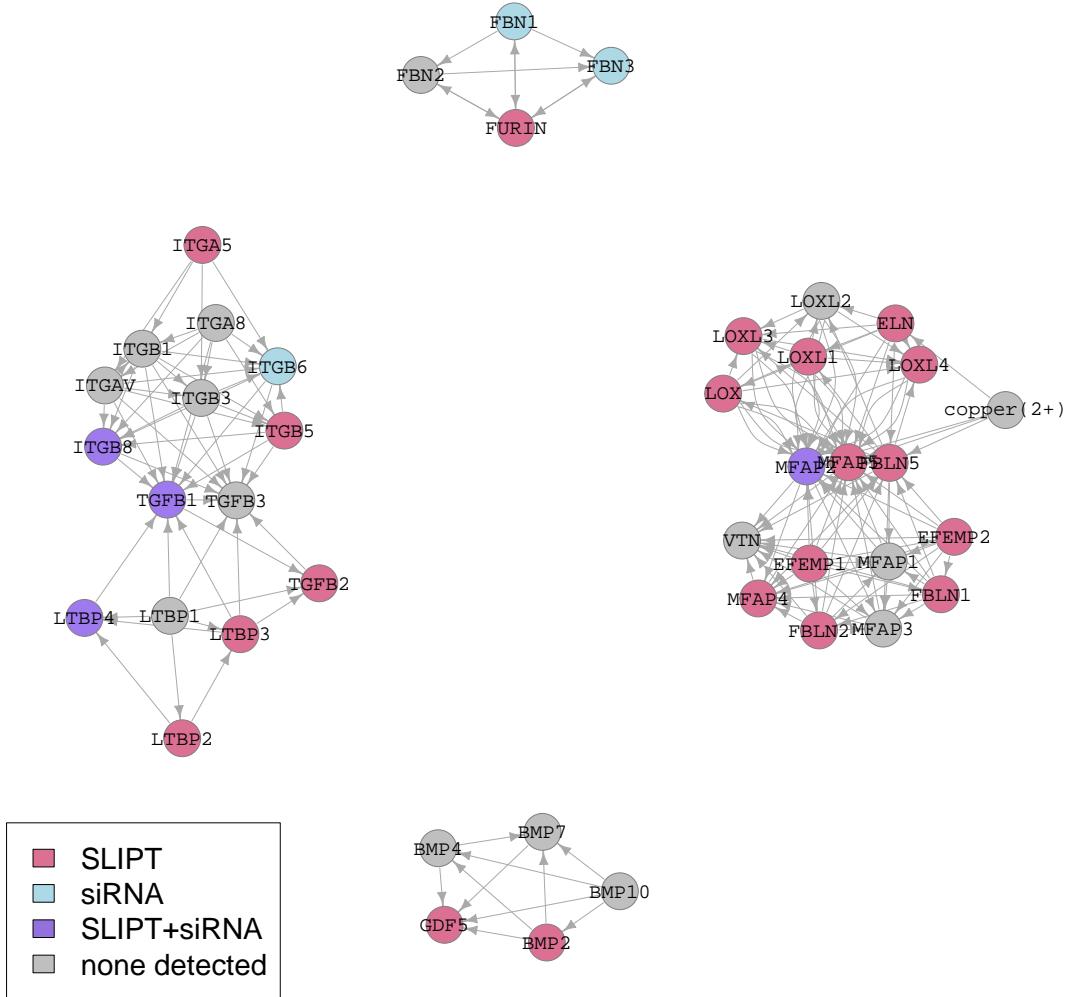


Figure 5.2: Synthetic Lethality in the Elastic Fibre Formation Pathway. The Reactome Elastic Fibre Formation pathway with synthetic lethal candidates coloured as shown in the legend.

be downstream of those detected by SLIPT, in addition to whether connectivity or centrality is higher for synthetic lethal candidates than other genes in the pathway.

Genes detected as synthetic lethal partners of *CDH1* by SLIPT or siRNA screening were also common in the Fibrin clot formation pathway (shown in Figure 5.3). This is consistent with the established pleiotropic role of *CDH1* in regulating fibrin clotting. It is also notable that the genes detected by either method appear to be highly connected such as *C1QBP*, *KNG1*, *F8*, *F10*, *F12*, *F13A*, and *PROC* (including many of the

coagulation factors). Synthetic lethal candidates also include *SERPINE2* and *PRCP*, which only affect downstream genes, in addition to *PROCR* and *VWF*, which are only affected by upstream genes.

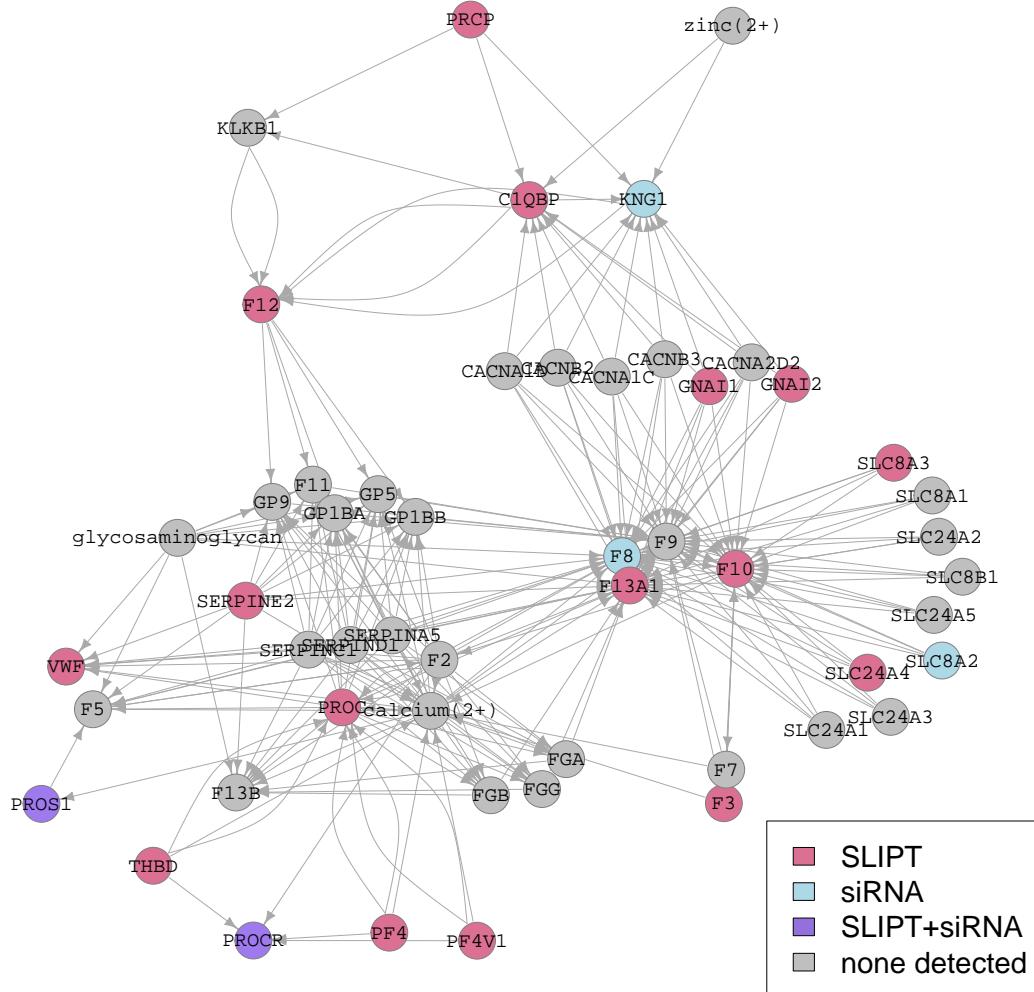


Figure 5.3: Synthetic Lethality in the Fibrin Clot Formation. The Reactome Fibrin Clot Formation pathway with synthetic lethal candidates coloured as shown in the legend.

Many of these genes are involved in the larger Extracellular Matrix pathway (shown in Appendix Figure F.3), including many of the synthetic lethal candidates discussed for elastic fibres. The number of SLIPT candidate genes outnumbers those identified by siRNA, as expected from an isolated cell model. However, the endocrine response genes (e.g., *TGFB1* and *LTBP4*) which are potentially artifacts of the cell line growth

process were replicated with SLIPT analysis in patient tumours (TCGA breast cancer data). There is also additional support for synthetic lethal genes (e.g., *ITGB2*, *MFAP2*, and *SPARC*) being highly connected networks hubs of the pathway. The complexity of the extracellular matrix pathway lends credence to the need for formal network analysis approaches to interpret the pathway structure of synthetic lethal candidates. Furthermore statistical approaches are needed to determine whether structural relationships are unlikely to be observed between synthetic lethal candidates by sampling error.

5.1.3 G Protein Coupled Receptors

G protein coupled receptor (GPCR) pathways are highly complex (as shown in Appendix Figures F.4 and F.5). Many of genes in these pathways were synthetic lethal candidates, detected by either SLIPT or siRNA screening, including genes frequently detected with both approaches, consistent with these pathways being supported by prior analyses (in Sections 4.2.5 and 4.2.5.1). Synthetic lethal candidates include the PDE and Calmodulin genes (as discussed in Section 5.1.3) in addition to others such as the regulators of G-protein signalling (RGS), chemokine receptors (CXCR), Janus kinase (JAK), and the Ras homolog family (RHO) genes. These are important regulatory signalling pathways necessary for cellular growth and cancer proliferation. Thus the GPCR pathways (and downstream PI3K/AKT signals) are a potentially actionable vulnerability against *CDH1* deficient cancers, particularly since many existing drug targets exist among these signalling pathways, some of which have been experimentally validated (Kelly *et al.*, 2017b; Telford *et al.*, 2015). However, the complexity of GPCR networks containing hundreds of genes requires the relationships between SLIPT and experimental candidates to be tested with a network based statistical approach, although statistically significant number of genes in GPCR pathways was detected by both approaches (in Sections 4.2.5 and 4.2.5.1).

5.1.4 Gene Regulation and Translation

While very few synthetic lethal genes were detected in translational pathways in an experimental screen against *CDH1* (Telford *et al.*, 2015), these were highly overrepresented in translational elongation (as shown in Appendix Figure F.6). These SLIPT genes include many ribosomal proteins and the regulatory “elongation factors” which may be subject to responses in the upstream signalling pathways. This observation lends support to the notion of pathway structure among synthetic lethal candidates detected by SLIPT in comparison with siRNA. The computational approach with SLIPT displays the ability to detect downstream genes in the core translational

processes which experimental screening did not identify. The experimental screening may similarly detect upstream regulatory genes less sensitive to inactivation, that is, genes that are less likely to be indiscriminately lethal to both genotypes at high doses of inactivation.

Many of these SLIPT candidate genes are also among the nonsense-mediated decay (NMD) pathway (shown in Appendix Figure F.7) or 3' untranslated region (UTR) mediated translational regulation (shown in Appendix Figure F.8). While genes in these pathways were also supported by experimental screening with siRNA, there was differences in which genes were detected within the pathway structures. In particular, *UPF1* was detected in the siRNA screen and is the focal downstream gene for the entire NMD pathway showing that (in this case) siRNA genes are downstream effectors of those detected by SLIPT. 3' UTR mediated translational regulation has a similar structure with two modules connected solely by *RPL13A*, giving an example of SLIPT candidate genes with high connectivity, although there were many ribosomal proteins detected by SLIPT. However, the detection of *EIF3K*, a regulatory elongation factor (not essential to ribosomal function) was replicated across SLIPT and siRNA screening, while the majority of the elongation factors were not detected by either approach. Regulatory genes, being more amenable to experimental validation, also support further investigation into pathway structure. The SLIPT candidates may support experimental candidates in biological pathways by detecting downstream genes, which may not be detectable by experimental screening with high dose inhibitors. This difference between the approaches may explain the greater number of SLIPT candidate partners of *CDH1* than those experimentally identified.

5.2 Network Analysis of Synthetic Lethal Genes

Genes detected as synthetic lethal partners of *CDH1* with the SLIPT computational approach and the siRNA screen (Telford *et al.*, 2015) were compared across network metrics in the example of the PI3K cascade pathway (where the genes differed considerably between synthetic lethal detection methods). These were used to test whether network metrics differed between groups of genes detected by either or both approaches. These analyses serve to test both whether synthetic lethal gene candidates had higher connectivity or importance in a network and whether either detection approach is biased towards genes with different network properties.

5.2.1 Gene Connectivity and Vertex Degree

Vertex degree (the number of connections) for each gene is a fundamental property of a network. The vast majority of genes had a relatively modest number of connections, each with only a few genes in the PI3K pathway (shown in Figure 5.4) having pathway relationships with a high number of genes, consistent with the scale-free property of biological networks (Barabási and Oltvai, 2004). There were few differences in the number of connections between gene groups (by synthetic lethal detection), although genes detected by siRNA included those with the fewest connections. The median connectivity of genes detected by both approaches was marginally higher.

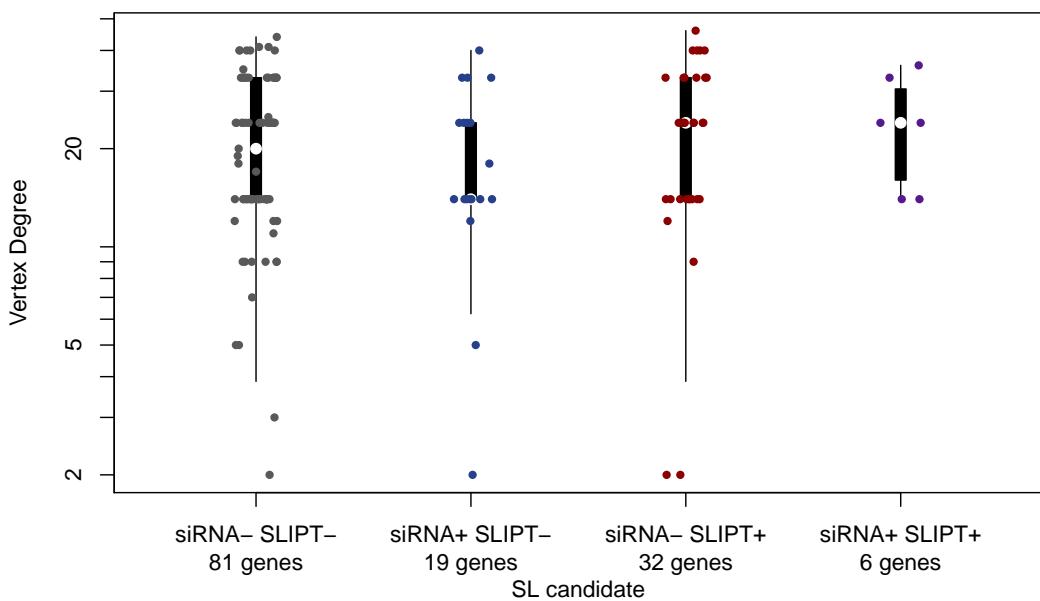


Figure 5.4: **Synthetic Lethality and Vertex Degree.** The number of connected genes (vertex degree) was compared (on a log-scale) across genes detected by SLIPT and siRNA screening in the Reactome PI3K cascade pathway. There were very few differences in vertex degree between the groups, although genes detected by siRNA included those with the fewest connections.

Table 5.1: Analysis of variance (ANOVA) for Synthetic Lethality and Vertex Degree

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	15	15.50	0.0134	0.9082
SLIPT	1	506	506.01	0.4378	0.5105
siRNA × SLIPT	1	0	0.05	0.0000	0.9947

Analysis of variance for vertex degree against synthetic lethal detection approaches (with an interaction term)

The results for the PI3K pathway were very similar when testing synthetic lethality against *CDH1* mutation (mtSLIPT). In this case, there is also indication that mtSLIPT-specific genes may have higher connectivity than those detected by siRNA screening (shown in Appendix Figure G.1).

However, these apparent differences in vertex degree may be due to fewer genes being detected by either approach. There was no statistically significant effect of either computational or experimental synthetic lethal detection method on vertex degree, as determined by ANOVA (shown by Table 5.1 and Appendix Table G.1). Thus synthetic lethal detection does not discriminate among genes by their connectivity in a pathway network, nor is either approach constrained to detecting highly connected genes. Both approaches have been demonstrated to detect genes with many and very few connections.

5.2.2 Gene Importance and Centrality

5.2.2.1 Information Centrality

Information centrality is a measure of the importance of nodes in a network by how vital they are to the transmission of information throughout the network. This applies well to biological pathways, particularly gene regulation and cell signalling. The nodes with the highest information centrality are not necessarily the most connected, as they may also include nodes that pass signals between highly connected network hubs. Information centrality therefore provides a distinct metric for the connectivity of a gene in a pathway, which has the added benefit of being directly related to the disruption of pathway function were it to be inactivated or removed.

Information centrality has also been suggested to indicate essentiality of genes or proteins (Kranthi *et al.*, 2013). The information centrality for each gene was computed across the entire Reactome network (as discussed in Appendix H). Reactome contains substrates and cofactors in addition to genes and proteins. In support of centrality

as a measure of essentiality or importance to the network, a number of nodes with the highest centrality (shown in Appendix Table H.1) were essential nutrients, including Mg^{2+} , Ca^{2+} , Zn^{2+} , and Fe.

Genes important in development of epithelial tissues and breast cancer were also detected with relatively high information centrality (as shown by the distribution across the Reactome network in Appendix Figure H.1). Interleukin 8 (encoded by *IL8*) is a chemokine important in epithelial cells, the innate immune system, and binding GPCRs. *GATA4* is an embryonic transcription factor involved in heart development, EMT, and has been shown to be recurrently mutated in breast cancer (TCGA, 2012). β -catenin (encoded by the proto-oncogene *CTNNB1*) is a regulatory protein which binds to E-cadherin, being involved in cell-cell adhesion and Wingless-related integration site (WNT) signalling. Together these show that information centrality identifies nodes of importance to biological functions in pathway networks, including those relevant to *CDH1* deficient breast cancers.

Within the PI3K pathway, genes detected by siRNA did not include those with lower centrality (shown in Figure 5.5), although the median information centrality across gene groups detected by either synthetic lethal approach did not differ. The genes with the highest information centrality included the synthetic candidates *PDE3B* (detected by SLIPT and siRNA) and *AKT2* (detected by SLIPT) which were markedly higher than most other genes in the pathway. The higher centrality of these genes is consistent with their known biological role in PI3K/AKT signalling and the pathway structure (shown in Figure 5.1). Other biomolecules with high centrality included the *RPS6KB1* and *RPTOR* genes, adenosine monophosphate (AMP), phosphatidylinositol (4,5)-bisphosphate (PIP_2), and phosphatidylinositol (3,4,5)-trisphosphate (PIP_3).

These findings were replicated (shown in Appendix Figure G.2) when testing synthetic lethality against *CDH1* mutation (mtSLIPT). The differences in network centrality between gene groups detected by either method were not statistically significant as determined by ANOVA (shown by Table 5.2 and Appendix Table G.2). Thus neither method was unable to detect synthetic lethal genes with particular centrality constraints, although they were also not detecting genes with higher centrality than expected by chance.

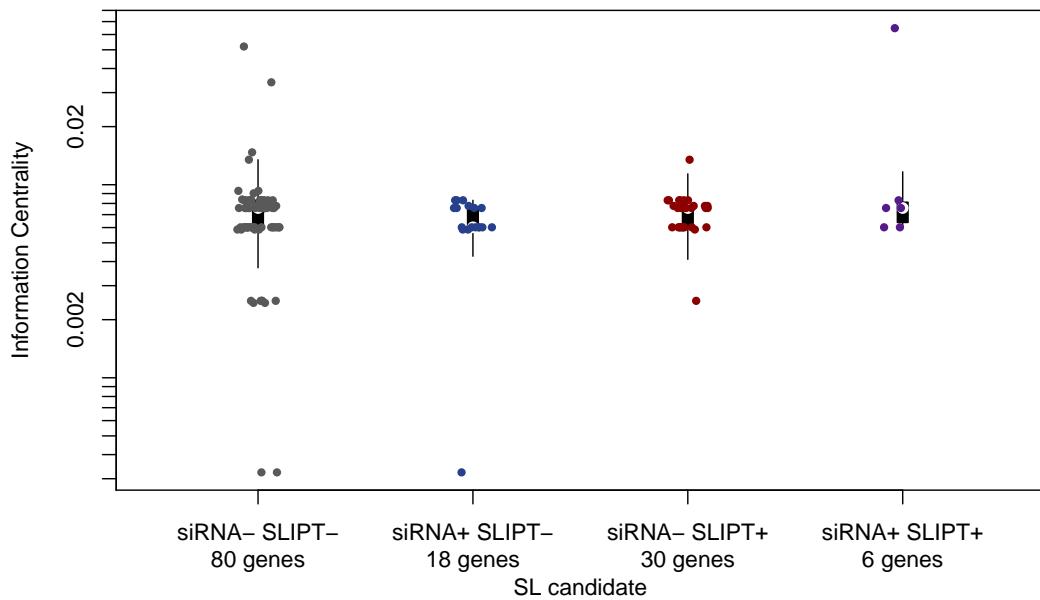


Figure 5.5: Synthetic Lethality and Centrality. The information centrality was compared (on a log-scale across genes detected by SLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by SLIPT or siRNA did not have higher connectivity than other genes. The gene with the highest centrality was detected by both approaches.

Table 5.2: ANOVA for Synthetic Lethality and Information Centrality

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.000256	0.0002561	0.1854	0.6682
SLIPT	1	0.003827	0.0038275	2.7717	0.1008
siRNA×SLIPT	1	0.000804	0.0008036	0.5820	0.4483

Analysis of variance for information centrality against synthetic lethal detection approaches (with an interaction term)

5.2.2.2 PageRank Centrality

PageRank centrality is another network analysis procedure to infer a hierarchy of gene importance from a network using connections and structure (Brin and Page, 1998). In contrast to the information centrality approach of removing nodes, PageRank uses the eigenvalue properties of the adjacency matrix to rank genes according to the number of connections and paths they are involved in.

This distinction is immediately clear within the PI3K pathway (shown in Figure 5.6), which differs considerably from the information centrality scores. Genes detected by SLIPT span the complete range of PageRank centrality values for this pathway, which was replicated when testing synthetic lethality against *CDH1* mutation (shown in Appendix Figure G.3). However, the genes detected by both SLIPT and siRNA screening have a higher median PageRank centrality, although the differences in PageRank centrality between these methods were not statistically significant as determined by ANOVA (shown by Table 5.3 and Appendix Table G.3).

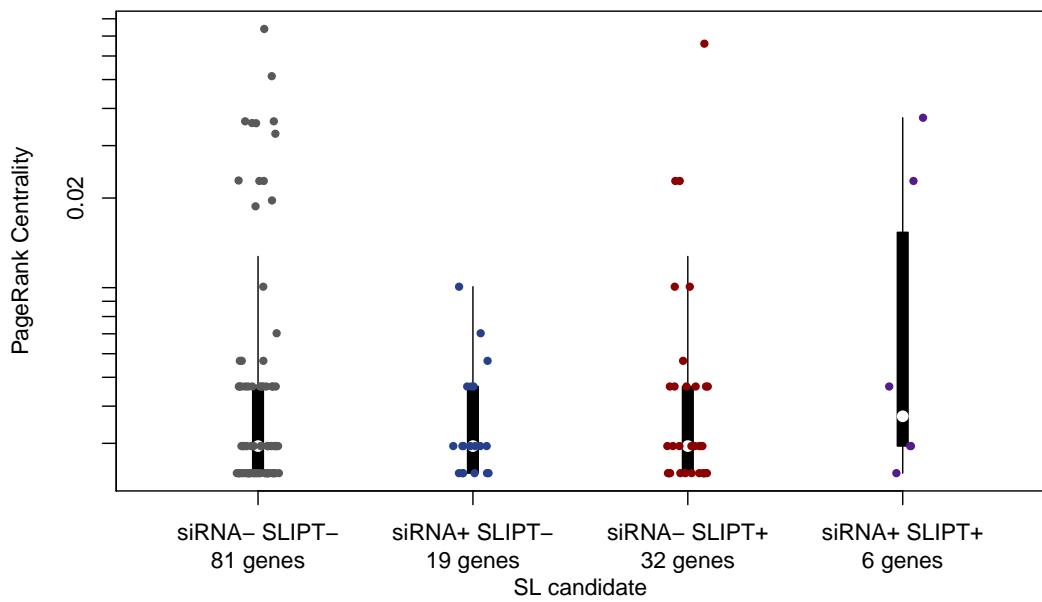


Figure 5.6: Synthetic Lethality and PageRank. The PageRank centrality was compared (on a log-scale across genes detected by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had a more restricted range of centrality values (which may be constrained experimental detection in a cell line model) than other genes not detected by either approach, although these groups also had fewer genes and a higher median.

Table 5.3: ANOVA for Synthetic Lethality and PageRank Centrality

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.0002038	2.0385×10^{-4}	1.1423	0.2892
SLIPT	1	0.0000208	2.0752×10^{-5}	0.1163	0.7342
siRNA \times SLIPT	1	0.0000137	1.3743×10^{-5}	0.0770	0.7823

Analysis of variance for PageRank centrality against synthetic lethal detection approaches (with an interaction term)

5.3 Relationships between Synthetic Lethal Genes

5.3.1 Hierarchical Pathway Structure

5.3.1.1 Contextual Hierarchy of PI3K

A contextual hierarchy of genes in the PI3K pathway was performed (as described in Section 3.4.1.2) to assign scores for their relative order in the pathway. In the case of PI3K (shown in Figure 5.7), this orders genes from the upstream genes, which respond to signals from extracellular stimuli, to the downstream genes which transmit these to the gene expression (translation) responses of the cell. The directionality of this pathway is evident in transmitting signals from the PI3K complex, via AKT, PDE, and mTOR to the ribosomal regulatory proteins. This hierarchical procedure enables testing whether the biological context of a gene in a pathway is relevant to detection as a synthetic lethal candidate by either computational SLIPT analysis or experimental siRNA screening.

5.3.1.2 Testing Contextual Hierarchy of Synthetic Lethal Genes

This pathway hierarchy in the PI3K cascade was tested for differences between genes detected across SLIPT and siRNA screening. The synthetic lethal candidates for *CDH1* detected by either method (as shown by Figure 5.8a) did not differ, each being distributed throughout the pathway. When adjusted for being more numerous, there was little indication that SLIPT candidate genes are more frequently upstream or downstream of siRNA candidate genes (as shown by Figure 5.8b) and were more frequent at moderate hierarchies which contained more genes. Synthetic lethal candidates from both methods were less frequently detected in the downstream effectors of the pathway (e.g., the mTOR complex), although core pathway genes (e.g., *AKT2* and *PDE3B*) were detectable as synthetic lethal candidates (as discussed for Figures 5.1 and 5.6).

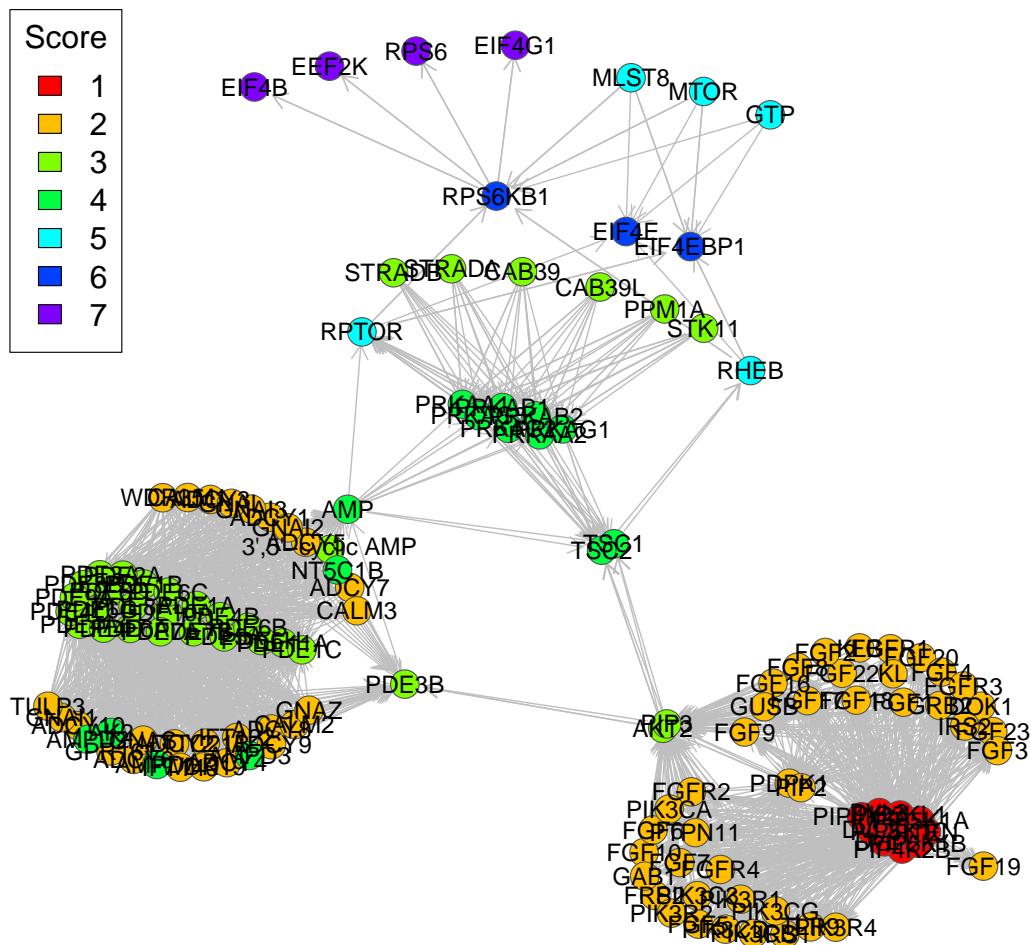
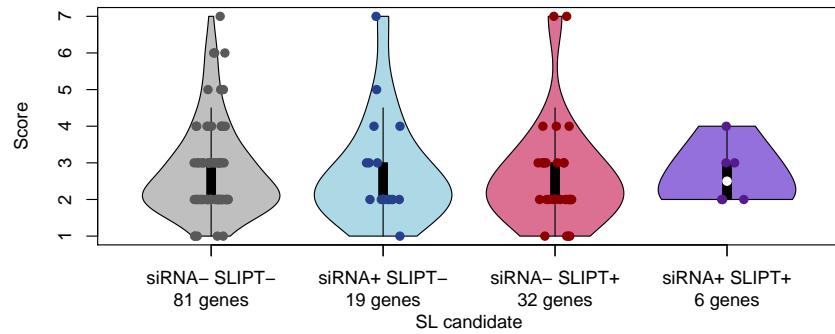


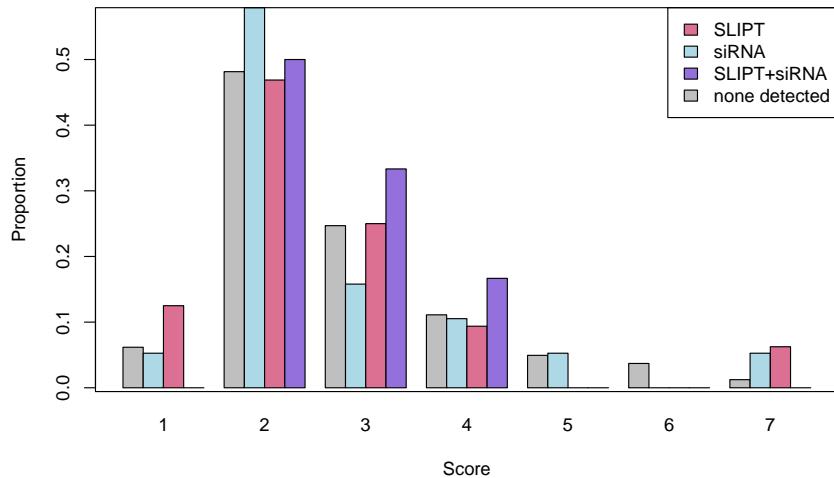
Figure 5.7: Hierarchical Structure of PI3K. A contextual score was used for ranking genes within the PI3K Cascade to demonstrate a pathway structure analysis to examine whether genes detected by either SLIPT or siRNA were more frequently upstream or downstream in the PI3K pathway.

Similarly, when testing synthetic lethality against *CDH1* mutation (mtSLIPT), the hierarchical score for the PI3K pathway did not differ between mtSLIPT-specific and siRNA-specific gene candidates (as shown by Appendix Figure I.1). The median among genes detected by both approaches was marginally elevated such that these genes may be further downstream in the pathway than other synthetic lethal candidate partners of *CDH1*. There were fewer genes overall with higher scores (shown in Appendix

Figure I.2). While these were more frequently detected by both SLIPT and siRNA, there was no significant effect variation in pathway hierarchy (shown by ANOVA in Table 5.4 and Appendix Table I.1) accounted for by SLIPT or siRNA detection in the PI3K pathway (as shown in Figure 5.1). Thus these hierarchical scores may be observed by sampling variation and there is no indication that SLIPT or siRNA detection differs



(a) Hierarchical Distance Score



(b) Proportion of Genes

Figure 5.8: Hierarchy Score in PI3K against Synthetic Lethality in PI3K. The hierarchical distance scores were similarly distributed across SLIPT and siRNA genes. The number of SLIPT and siRNA genes against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

along the direction of the pathway. Genes detected by either method are no more or less common among upstream or downstream of the pathway.

Table 5.4: ANOVA for Synthetic Lethality and PI3K Hierarchy

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.001	0.00066	0.0004	0.9842
SLIPT	1	0.456	0.45605	0.2740	0.6016
siRNA×SLIPT	1	0.019	0.01878	0.0113	0.9156

Analysis of variance for PI3K hierarchy score against synthetic lethal detection approaches (with an interaction term)

[remove this paragraph and Figures 5.9 and I.3?]

Furthermore the pathway hierarchical scores did not exhibit different more or less SLIPT than siRNA genes above or below the given threshold. Since the ideal threshold to detect pathway structure is unclear, an exploratory analysis was performed, with χ^2 -test for the SLIPT or siRNA candidate genes upstream or downstream of each gene. It is unsurprising that these χ^2 tests were highest when the gene used as a threshold was in the middle of the pathway (as shown in Figure 5.9). However, there was no statistically significant support for pathway structure by this approach, as none of the χ^2 values were high enough to detect pathway structure between SLIPT and siRNA gene candidates. Nor was structure detectable for mtSLIPT testing synthetic lethality against *CDH1* mutation (as shown in Appendix Figure I.3).

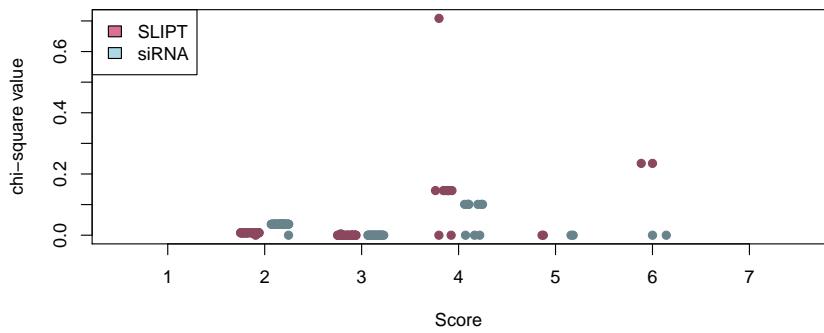


Figure 5.9: Structure of Synthetic Lethality in PI3K. The number of SLIPT and siRNA genes upstream or downstream of each gene in the Reactome PI3K pathway were tested (by the χ^2 -test). These are plotted as a split jitter stripchart against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

5.3.2 Upstream or Downstream Synthetic Lethality

This approach does not ascertain whether SLIPT and siRNA candidate partners of *CDH1* are upstream or downstream of one and other within a pathway such as the PI3K cascade. The hierarchical approach is designed to detect differences in pathway location between gene groups. An alternative pathway structure method has been devised to use network structures to identify directional relationships between individual SLIPT and siRNA genes. This pathway structure methodology will be applied (as described in Section 3.4.1) to detect the direction of shortest paths between SLIPT and siRNA gene candidates. This will be used to demonstrate the methodology on the PI3K pathway, to develop a statistical test for pathway structure between between SLIPT and siRNA gene candidate using resampling (as described in Section 3.4.1.1), and to apply this test for pathway structure among synthetic lethal gene candidates to the pathways identified in Chapter 4 and discussed in Section 5.1.

5.3.2.1 Measuring Structure of Candidates within PI3K

Shortest paths in a pathway network were used to devise a strategy to detect pathway structure between SLIPT and siRNA gene candidate partners of *CDH1* (as described in Section 3.4.1). Thus we can determine whether individual SLIPT genes have upstream or downstream siRNA candidates (scored as “up” or “down” events respectively). This

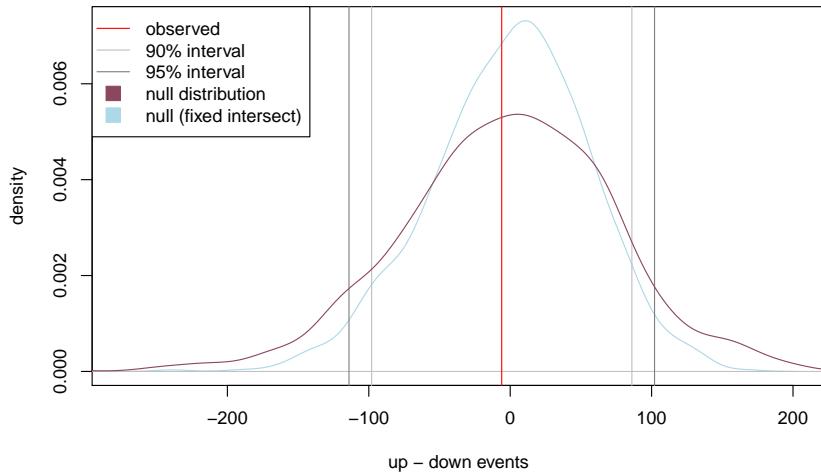


Figure 5.10: Structure of Synthetic Lethality Resampling in PI3K. A null distribution with 10,000 iterations of the number of siRNA genes upstream or downstream of SLIPT genes (depicted as the difference of these) in the PI3K pathway. To assess significance, the observed events (with shortest paths) were compared to the 90% and 95% intervals for the null distribution (shown in violet). Genes detected by both methods were fixed to the same number as observed for the alternative null distribution (shown in blue), although the observed number of events (red) was not significant in either case. In both cases, these genes detected by both approaches were included in computing the number of shortest paths (in either direction) between SLIPT and siRNA genes.

procedure enables the detection of directional relationships between SLIPT and siRNA gene candidates (in contrast to the hierarchical approach).

The total number of gene candidate pairs in either direction can be compared within a pathway network to assess the overall directional relationships in a pathway. This directionality is detectable by the difference between the number of SLIPT candidate genes with upstream and downstream siRNA gene partners. However, this measure alone is not sufficient to determine whether there is evidence of pathway structure between SLIPT and siRNA gene candidate partners of *CDH1* in a pathway network. Nevertheless, it does serve to measure the magnitude (and direction) of the consensus of directional relationships (upstream and downstream) between SLIPT and siRNA gene candidate partners. This measure of pathway structure can be used for testing for statistical significance of pathway structure by resampling, using a permutation procedure to test whether these relationships are detectable among randomly selected

gene groups rather than the detected SLIPT and siRNA gene candidate partners (as described in Sections 2.3.6 and 3.4.1.1).

This resampling procedure was performed for the PI3K network to generate a null distribution for the difference in the number of “up events” and “down events” for this pathway (as shown in Figure 5.1). Resampling yields a distribution to detect whether genes detected by SLIPT had significantly more upstream or downstream siRNA candidates. While there was modest indication that siRNA genes were downstream of SLIPT candidate genes, resampling for the PI3K pathway (as shown in Figure 5.10) did not detect a significant number of siRNA genes upstream or downstream.

In contrast, when testing synthetic lethality against *CDH1* mutation (mtSLIPT) there was modest indication that siRNA genes were upstream of SLIPT candidate genes. However, resampling (as shown in Appendix Figure I.4) was also unable to detect a significant number of siRNA genes upstream or downstream of mtSLIPT candidates. Neither fixing the number of genes detected by both approaches (as shown by the blue line in Figure 5.10 and Appendix Figure I.4) nor excluding these jointly detected genes altered the findings of this approach. These genes were included in the analysis because they can disproportionately count towards siRNA genes being upstream (or downstream) of SLIPT genes as they may still have different proportions of gene detected by either approach upstream (or downstream) of them. Furthermore, expanding the range of shortest paths to consider links in related pathways (using the “metapathways” constructed in Section 2.4.3) also had little effect on the null distribution generated, despite increasing the computational demands of the procedure.

5.3.2.2 Resampling for Synthetic Lethal Pathway Structure

The permutation procedure (as described in Section 3.4.1.1) that was performed in Section 5.3.2.1 for the PI3K cascade was also applied to other pathways identified in Chapter 4 and discussed in Section 5.1. These include extracellular matrix (with constituent elastic fibre and fibrin pathways), cell signalling (by PI3K/AKT and GCPRs), and translational pathways (with NMD and 3'UTR regulation). The resampling results across these pathways (as shown in Table 5.5) had limited support for association between pathway structure and detection of synthetic lethal genes, with the majority of these being non-significant as shown for PI3K (in Appendix Figure I.4). However, the distribution for these pathways will differ depending on their structure, the number of genes they consist of, and the proportion of synthetic lethal candidates among them (including a higher frequency of genes detected by both methods for the pathways identified in Section 4.2.5.1). This resampling is an appropriate procedure to use to

detect structural relationships across pathways as it does not assume an underlying test statistic distribution.

Pathway structure was supported for the NMD pathway (which is consistent with siRNA being downstream in Appendix Figure F.7). However, this observation rests upon a single gene and was not replicated when testing synthetic lethality (mtSLIPT) against *CDH1* mutation (as shown in Appendix Table I.2) nor was it supported by the related 3'UTR regulation and translational elongation pathways.

Table 5.5: Resampling for pathway structure of synthetic lethal detection methods

Pathway	Graph		States		Observed				Permutation p-value	
	Nodes	Edges	SLIPT	siRNA	Up	Down	Up–Down	Up/Down	Up–Down	Down–Up
PI3K Cascade	138	1495	38	25	122	128	-6	0.953	0.5326	0.4606
PI3K/AKT Signalling in Cancer	275	12882	98	44	779	679	100	1.147	0.3255	0.6734
G_{αi} Signalling	292	22003	95	58	836	1546	-710	0.541	0.9971	0.0029
GPCR downstream	1270	142071	312	160	9755	9261	494	1.053	0.3692	0.6305
Elastic fibre formation	42	175	24	7	1	2	-1	0.500	0.5461	0.3865
Extracellular matrix	299	3677	127	29	547	455	92	1.202	0.3351	0.6636
Formation of Fibrin	52	243	18	5	12	17	-5	0.706	0.6198	0.3564
Nonsense-Mediated Decay	103	102	74	2	0	74	-74	0	1.0000	< 0.0001
3' -UTR-mediated translational regulation	107	2860	77	1	0	0	0		0.4902	0.5027
Eukaryotic Translation Elongation	92	3746	76	0	0	0	0		0.4943	0.4933

Pathways in the Reactome network tested for structural relationships between SLIPT and siRNA genes by resampling. The raw p-value (computed without adjusting for multiple comparisons over pathways) is given for the difference in upstream and downstream paths from SLIPT to siRNA gene candidate partners of *CDH1* with significant pathways highlighted in bold. Sampling was performed only in the target pathway and shortest paths were computed within it. Loops or paths in either direction that could not be resolved were excluded from the analysis. The gene detected by both SLIPT and siRNA (or resampling for them) were included in the analysis and the number of these were fixed to the number observed.

There does not appear to be a consensus on the directionality of SLIPT and siRNA candidates across pathways as distinct pathways showed stronger tendency for siRNA genes to be either upstream or downstream. Even related pathways such as PI3K and PI3K/AKT signalling showed directional events in opposite directions. The strongest pathway (among those tested) with support for directional pathways structure is G_{αi} signalling which showed significant downstream siRNA genes for both SLIPT and mt-SLIPT from a large number of shortest paths (in Table 5.5 and Appendix Table I.2). This would indicate that SLIPT detects upstream regulators of genes experimentally validated by siRNA. However, these results are borderline significant (with raw permutation p-values) and are unlikely to be detected after adjusting for multiple comparisons across the 10 pathways presented here (nor in the 1652 Reactome pathways used previously in Chapter 4).

Therefore, there is insufficient evidence to determine whether there is pathway structure, gene detected upstream or downstream by either method, between the SLIPT and

siRNA candidates in many of the synthetic lethal pathways (identified in Chapter 4). In particular, directional structure among synthetic lethal candidates for *CDH1* was not strongly supported in signalling pathways upon which the rationale for pathway structure hypotheses were based on. Despite the design of a robust resampling approach to test relationships between gene groups, this did not detect many structural relationships between SLIPT and siRNA gene candidates, although it may apply more broadly to gene networks. Furthermore, the pathway relationships are unlikely to be statistically supported by resampling when testing across the search space of Reactome pathways and adjusting for multiple comparisons. While there is statistically significant over-representation of many of these pathways in genes detected by both SLIPT and siRNA (as described in Chapter 4), these did not consistently show pathway structure. Furthermore, pathway structure did not account for the discrepancy between SLIPT and siRNA gene candidates which did not significantly intersect such as the PI3K cascade.

5.4 Discussion

These investigations used a functional pathway network that encapsulates protein complexes and functional modules. The Reactome network (Croft *et al.*, 2014) uses curated, experimentally identified pathways to determine relationships between genes and does not have the limitation of relying solely on protein binding or text-mining which are prone to false positives. While it is not documented whether these relationships are activating or inhibitory, the Reactome network (Croft *et al.*, 2014) is sufficient to test pathway relationships with directional information.

Synthetic lethal genes and pathways (for *CDH1* loss in cancer) were identified across gene expression and mutation datasets in Chapter 4. These pathway structure investigations extend those investigations into synthetic lethal gene candidates including exploring the discrepancy between SLIPT and siRNA candidate genes in a pathway such as PI3K in which they did not significantly intersect. Pathways with replicated synthetic lethal genes across these detection methods, breast and stomach cancer data, and patient and cell line data were also investigated including pathways from the extra-cellular microenvironment to core translational pathways and the signalling pathways between them.

Synthetic lethal gene candidates in the context of pathway structures can also be interpreted to provide additional mechanisms and support for belonging to a synthetic lethal pathway. Gene candidates with known mechanisms are ideal for triage of targets

specific to *CDH1* deficient tumours and for further experimental validation in preclinical models. This chapter presents computational methods to use pathway structure in an attempt to detect genes with importance in a pathway and reconcile the differences between SLIPT and siRNA candidate genes with pathway relationships (e.g., one group being downstream of the other).

Many genes were detected by either method and the differences between the computational and experimental screening approaches could feasibly lead to differences in which genes within a synthetic lethal pathway are identified. Genes detected by synthetic lethal detection strategies included those of biological importance within synthetic lethal pathways, those which are actionable drug targets, and those with functional implications for the biological growth mechanisms or vulnerabilities of *CDH1* deficient tumours. It appeared that genes detected by both approaches were highly connected (or of importance) in the network structure or some pathways and that there may be some structure with SLIPT and siRNA upstream or downstream of each other. However, the complexity of biological pathways meant that relationships between gene candidates were difficult to discern without formal mathematical and computational approaches and thus these were used to analyse large biological networks.

Network analysis techniques were therefore applied to formalise and quantify the connectivity and importance (centrality) of genes within pathways (using PI3K as an example). However, these network techniques were unable to identify distinct differences in the network properties of genes detected as synthetic lethal candidates by computational or experimental methods. These network metrics support the application of synthetic detection across pathways (and the findings using pathways as gene sets in Chapter 4) as neither synthetic lethal detection approach was biased towards genes of higher importance or connectivity and neither approach was insensitive to genes of lower importance or connectivity. SLIPT is therefore not biased towards genes with more crucial role in the pathway as inferred by pathway connectivity and centrality measures and detects genes irrespective of pathway structure.

Similarly, a network hierarchy based on biological context (ordered from receiving extracellular stimuli to affecting downstream gene expression and cell growth) was devised to test whether PI3K genes of a particular upstream or downstream level were more frequently detected as synthetic lethal candidates. However, this approach was unable to ascertain whether genes detected by either method were further upstream or downstream in the pathway and there was no statistical evidence that either method differed in which levels of this structure were detected.

A measure of pathway structure between individual SLIPT and siRNA genes within a pathway was also devised using the direction of shortest paths in a directed graph structure. This is amenable to detecting the consensus directionality of the pathway across pairs of genes detected by either method. The pathway structure methodology developed here is generally applicable to comparison of node groups (allowing overlapping) including genes in biological pathways and their detection by different methodologies. While the pathway structure measure alone is not able to detect structural relationships between gene groups (e.g., SLIPT and siRNA gene candidates), it is amenable to resampling to determine whether these relationships are statistically significant.

5.5 Summary

Together these analyses of biological pathways, network metrics, and statistical procedures devised specifically for this purpose were applied to Reactome pathway structures to test whether structural relationships exist between synthetic lethal candidates. Of particular interest was whether these relationships relate to the differences between the computational (SLIPT) and experimental (siRNA) synthetic lethal candidate partners of *CDH1* (in the pathways discussed in Chapter 4).

While biologically relevant relationships were observed in specific pathways, there were few detectable structural relationships between SLIPT and siRNA gene candidates. These candidates did not exhibit significant differences in network connectivity or centrality measures. Network analyses were also unable to ascertain whether the candidates detected by either method stratified into upstream and downstream genes on the pathway and they likely do not.

A statistical resampling procedure was applied to shortest path analysis to test whether pairs of SLIPT and siRNA gene candidates were more likely to be upstream or downstream of each other. This approach detected very few structural relationships in the synthetic lethal pathways identified in Chapter 4. Overall, support for pathway structure between SLIPT and siRNA gene candidates is weak and the direction is inconsistent between pathways. Therefore pathway structure does not account for the differences between the SLIPT and siRNA gene candidates, although this does support the validity of gene set analyses in Chapter 4 and the synthetic lethal pathways identified.

Furthermore, the resampling procedure demonstrated in this chapter is more widely applicable to gene states in network structures and may be of further utility in the anal-

ysis of biological pathways or networks. This approach was able to quantify structural relationships that were otherwise difficult to interpret and to conclusively exclude many potential relationships. In this respect, the network resampling methodology may also be applicable to triage of experimental validation.

Chapter 6

Simulation and Modeling of Synthetic Lethal Pathways

Simulation and modelling of synthetic lethality in gene expression is revisited in greater detail in this chapter, building upon the results provided (in Section 3.3) to support the use of SLIPT. 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 Comparing 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 enables 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 ??.

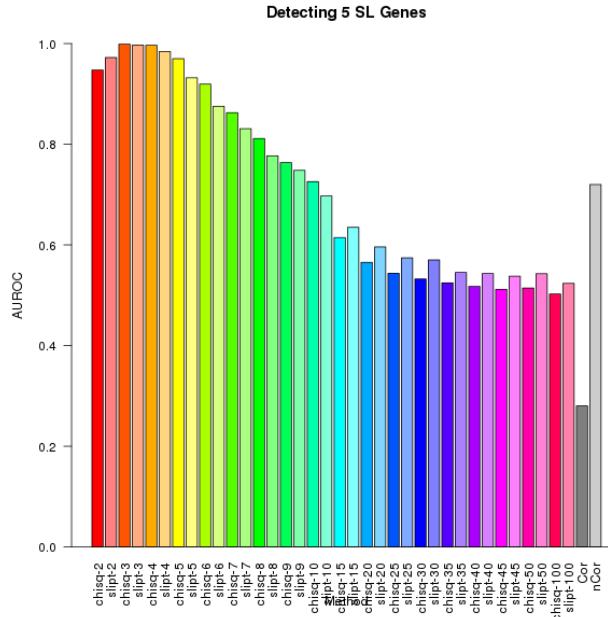
6.1.1 Performance of SLIPT and χ^2 across Quantiles

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 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.

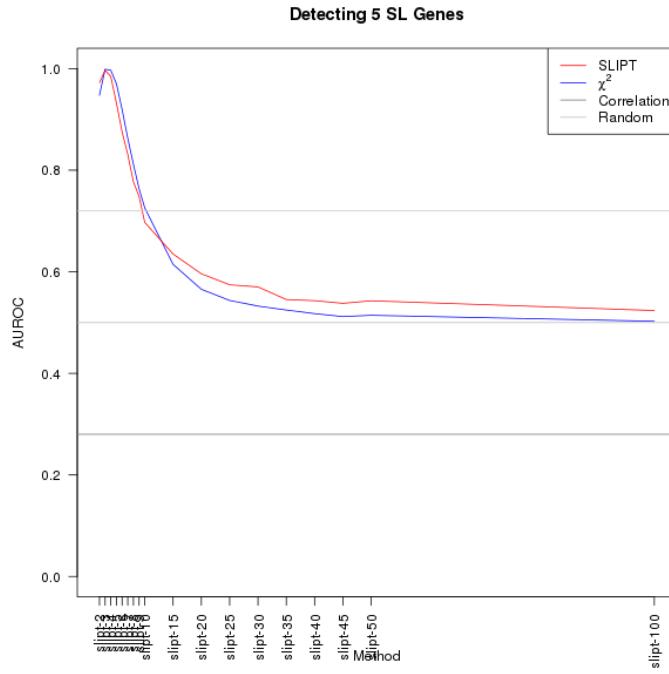
This enables testing of the threshold for lowly expressing genes which is most able to distinguish synthetic lethal genes, even with higher-order synthetic lethal interactions (as discussed in Section 3.2.1). Both SLIPT with the directional criteria for synthetic lethality and significance of the equivalent 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 fo 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



(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 in axis labels. The barplot uses the same hues for each quantile (grey for correlation) and darker for χ^2 (and positive correlation). The line plot 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 .

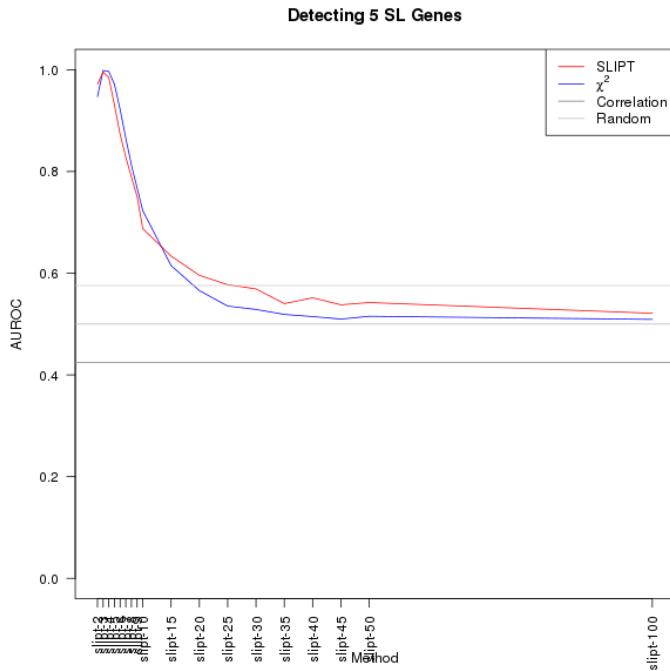


Figure 6.2: Performance of χ^2 and SLIPT across quantiles with more genes. Synthetic lethal detection (of 5 genes in 20,000) with quantiles as in axis labels. The line plot is coloured according to the legend. As for simulations with fewer genes, SLIPT and χ^2 perform similarly, peaking at $1/3$ -quantiles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

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 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

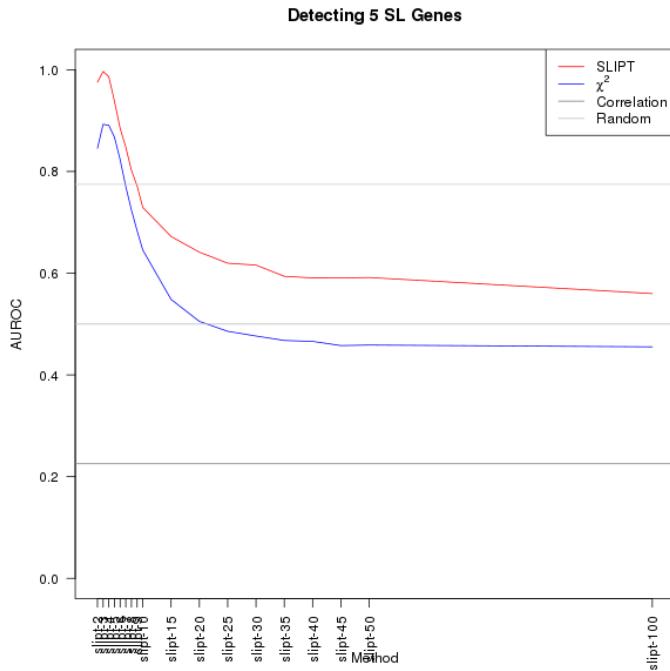


Figure 6.3: **Performance of χ^2 and SLIPT across quantiles with query correlation.** Synthetic lethal detection (of 5 genes in 100 including 5 query correlated) with quantiles as in axis labels. The line plot is coloured according to the legend. SLIPT performs consistently higher than χ^2 due to higher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

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).

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 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 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.

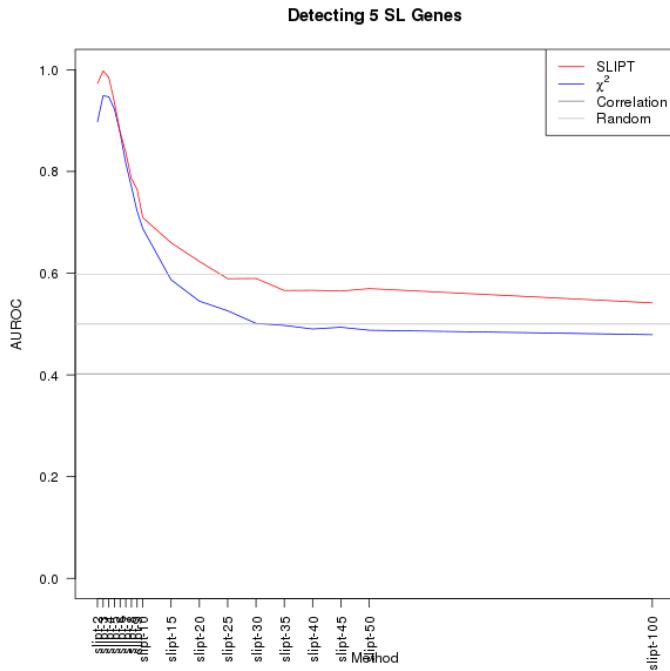


Figure 6.4: **Performance of χ^2 and SLIPT across quantiles with query correlation and more.** Synthetic lethal detection (of 5 genes in 20,000 including 1000 query correlated) with quantiles as in axis labels. The line plot is coloured according to the legend. SLIPT performs consistently higher than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

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 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 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.

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).

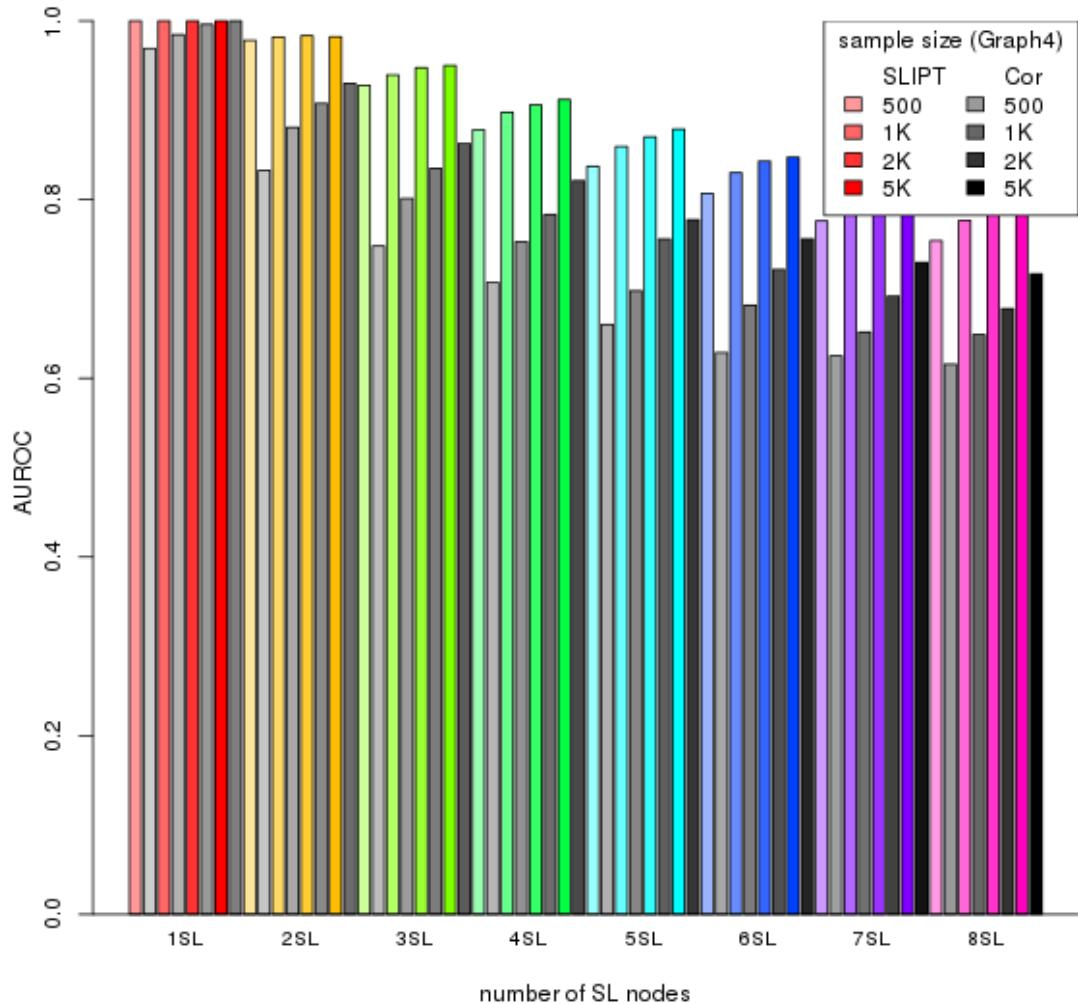


Figure 6.5: **Performance of negative correlation and SLIPT.** Synthetic lethal detection with SLIPT was compared to negative (Pearson) correlation across parameters. SLIPT consistently outperformed correlation. Both approaches had lower performance for more synthetic lethal partners and lower sample sizes.

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 genome 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

Simulations of synthetic lethality in Section 3.3 included correlated blocks of genes as a rudimentary model of pathway structure and co-regulated genes. Here the simulation procedure was expanded to account for more complex graph structures by sampling from multivariate normal distributions with correlation structure derived from graph structures (as described in Section 3.4.2). This approach enables simulation of synthetic

lethal pathways with known correlation structure and known partners (of a gene not in the pathway) and evaluation of the performance of SLIPT 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 will be tested both in artificial constructed networks to evaluate the effect of pathway structure on synthetic lethal detection, including large biologically feasible pathways to test whether SLIPT 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 enables 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) will be 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 will be 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 will be compared to the PI3K and $G_{\alpha i}$ signalling pathways derived from Reactome will be used for simulation of pathway structures of biological complexity (as shown in Figures 5.1 and F.4).

6.2.1 Performance over a Graph Structure

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. To start with, 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 small simple networks (as shown in Figures 6.7 and K.1).

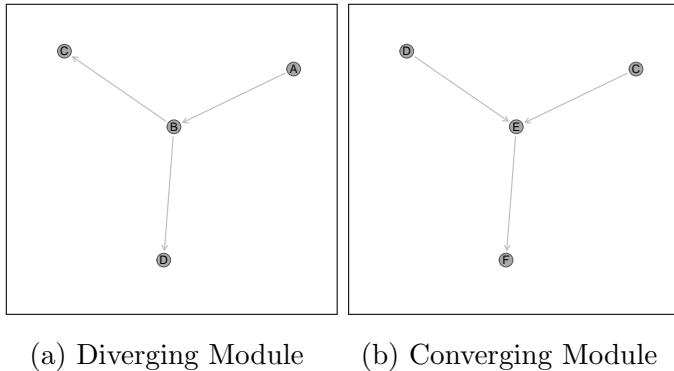


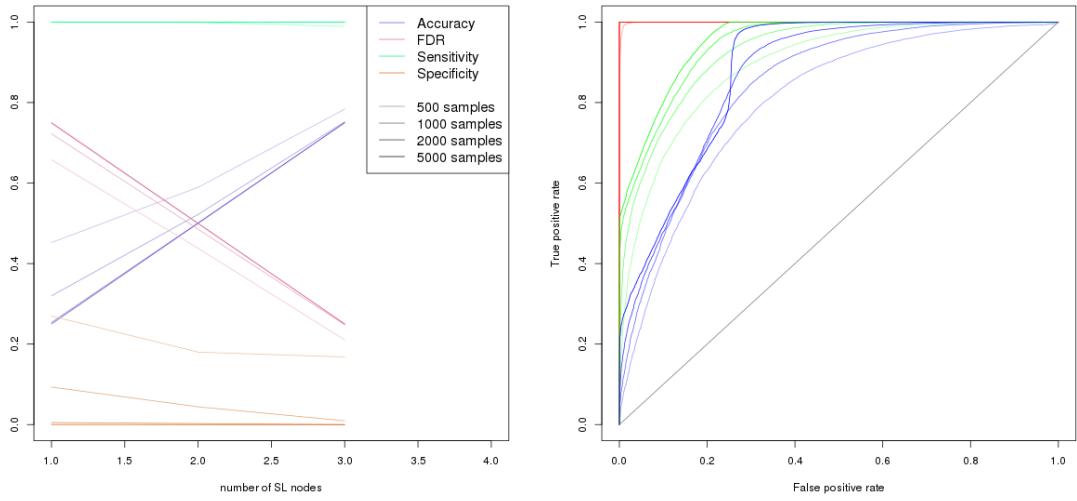
Figure 6.6: Simple graph structures. A simple graph structures used to demonstrate the simulation procedure. These are examples of a pathway diverging or converging respectively which enables testing the importance of direction in pathway structures. These are used with both activating and inhibiting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

As previously observed (in Section 3.3), performance declines with higher numbers of synthetic lethal genes and lower sample sizes. However, the sensitivity of SLIPT is high with conventional p-value thresholds (adjusted by FDR). Thus 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. Note that specificity cannot be tested if all potential partner genes are synthetic lethal which limits the number of synthetic lethal genes which can be tested.

These results are particularly consistent between the pathway modules of diverging (in Figure 6.8a) and converging signals (in Figure 6.8b), with the AUROC performance and underlying curves being strikingly similar between these graph structures (as shown in Figures 6.7 and K.1). This indicates that the performance of SLIPT is not perturbed by pathway structure, in particular the direction of pathway relationships, as these graph structures also demonstrate pathways in opposite direction. In a direct comparison (shown in Figure 6.8c), the performance of simulations in these simple graphs does not differ across parameter values and therefore SLIPT is robust to pathway direction.

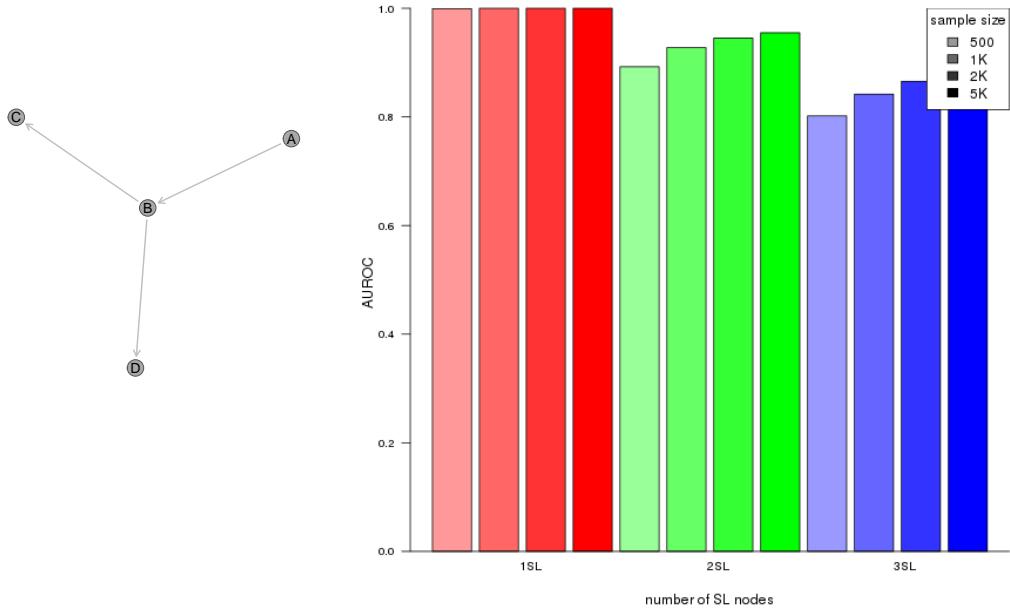
6.2.1.2 Constructed Graph Structures

A more complex graph structure was used to test the performance of detecting synthetic lethal partners with SLIPT in simulated expression data with pathway correlation



(a) Statistical evaluation

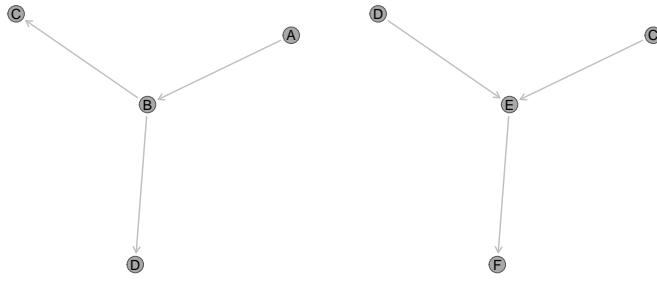
(b) Receiver operating characteristic



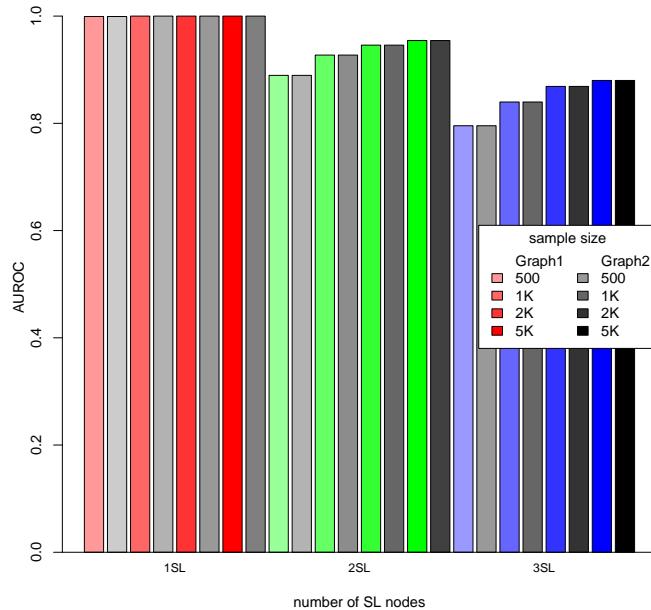
(c) Graph Structure

(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.



(a) Diverging Module (b) Converging Module

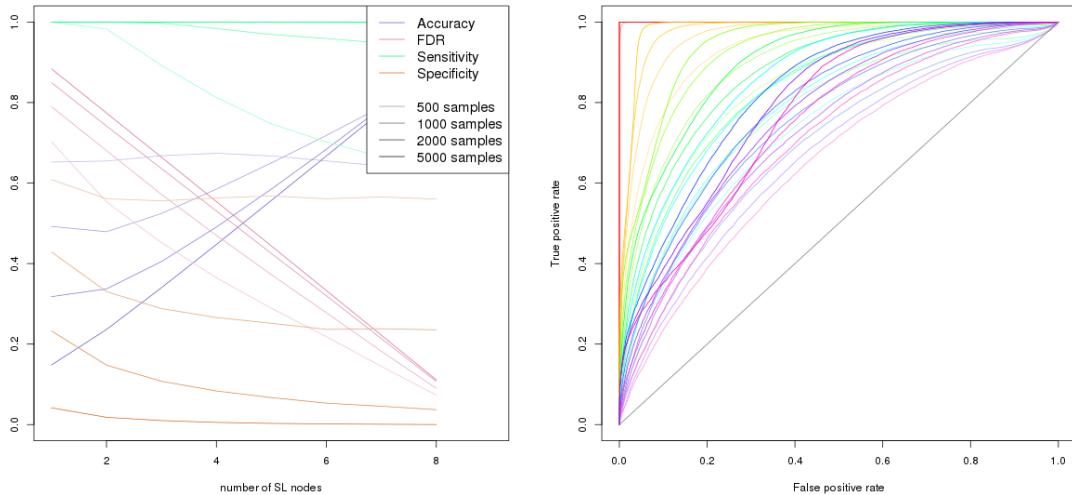


(c) Performance Between Graph Structures

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.

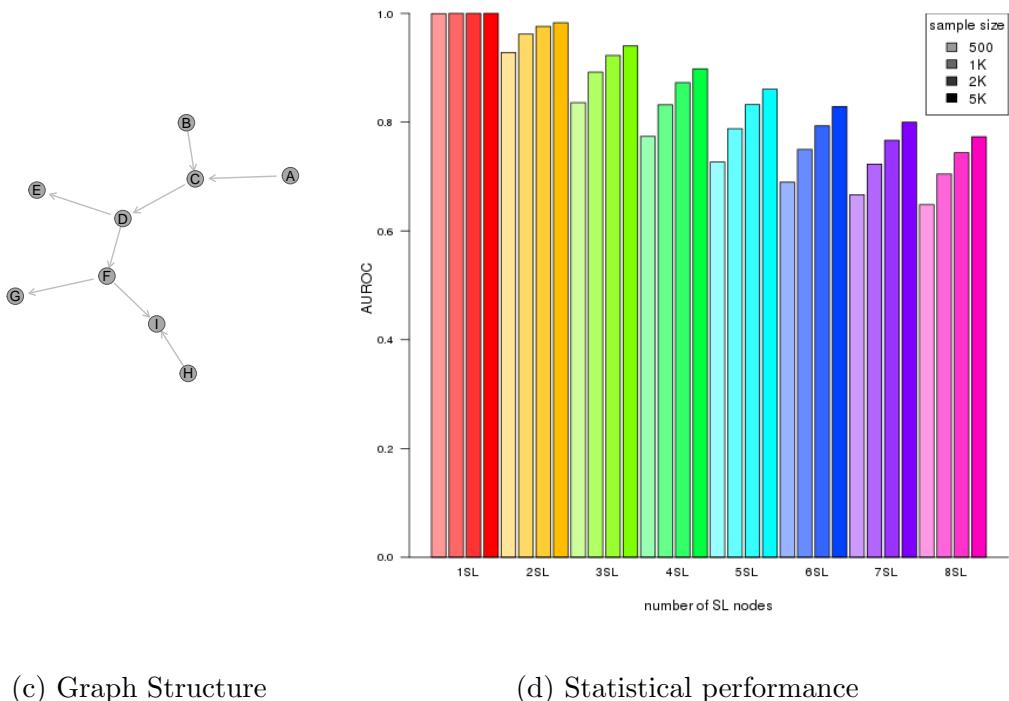
structures. For a simple chain of gene representing a pathway (shown in Figure 6.9), the above findings were generally replicated. Performance was high across parameter values in small networks, with similar decreases in higher numbers of synthetic lethal genes to detect and lower sample size.

When detecting synthetic lethal genes with SLIPT using adjusted (FDR) p-value thresholds, the performance differences can be largely attributed to changes in speci-



(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.

ficity as the small numbers of synthetic lethal genes produce 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.

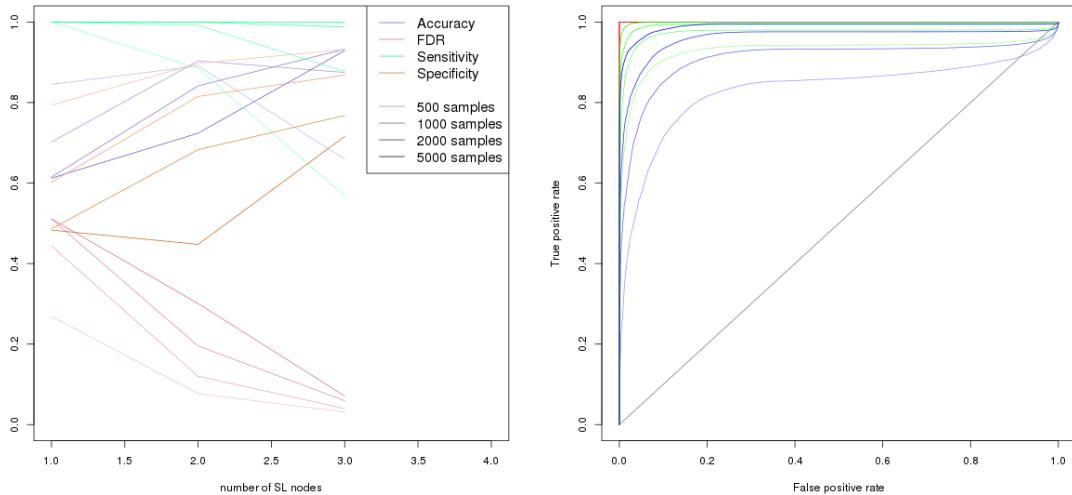
However, an artifact of these small test cases is the skewed ROC curves (as discussed in Section 6.2.1.1) which may be related to the low number of non-synthetic lethal genes to identify as true negatives, affecting the accuracy of specificity. This is unlikely to occur in large expression datasets with many non synthetic lethal genes, as shown previously (in Section 3.3) and 6.2.1.1) in simulations of graphs structures in larger datasets (in Section 6.2.4). This does not occur in larger, more complex graphs structures, even with modest total numbers of genes and high correlations (as shown in Section 6.3).

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).

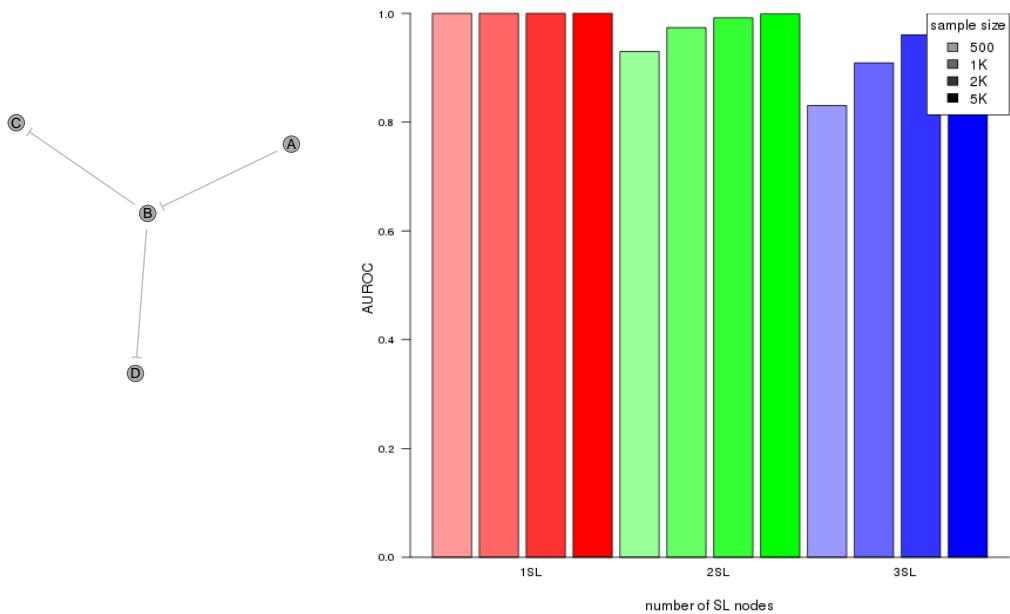
The ROC curves for an inhibiting graph also show consistently high specificity irrespective of detection threshold with only the upper extreme of the curve exhibiting a skew below random performance (in Figure 6.10). Nevertheless, the AUROC values show a high performance across parameter values, particularly 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



(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.

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).

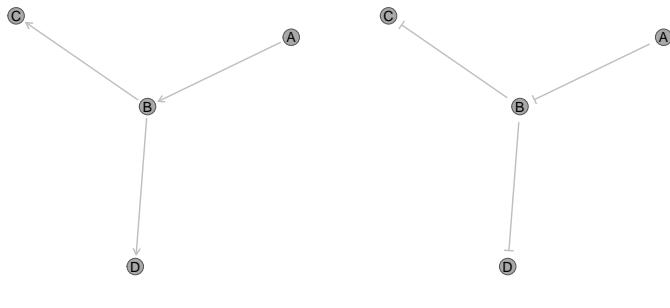
Detection of synthetic lethality by SLIPT in simulated data with inhibiting relationships outperforms simulations with activating relationships in the same graph structure (as shown in Figure 6.11). Thus SLIPT is robust in gene expression datasets with inverse correlations and performs well in them, 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 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.

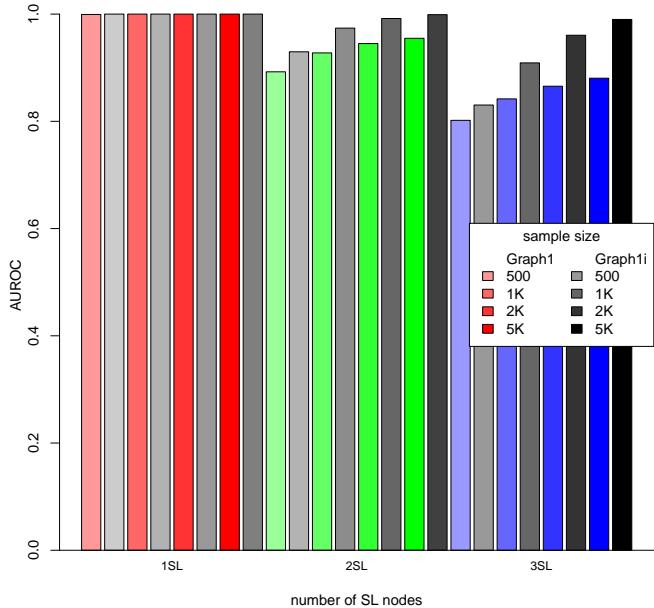
These still appear to have desirably high sensitivity, high accuracy, and low false positive rate for detecting more synthetic lethal genes, despite poor specificity. The ROC curves are particularly skewed for high proportions of the network being synthetic lethal and may stem from low numbers of true 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



(a) Activating Graph

(b) Inhibiting Graph

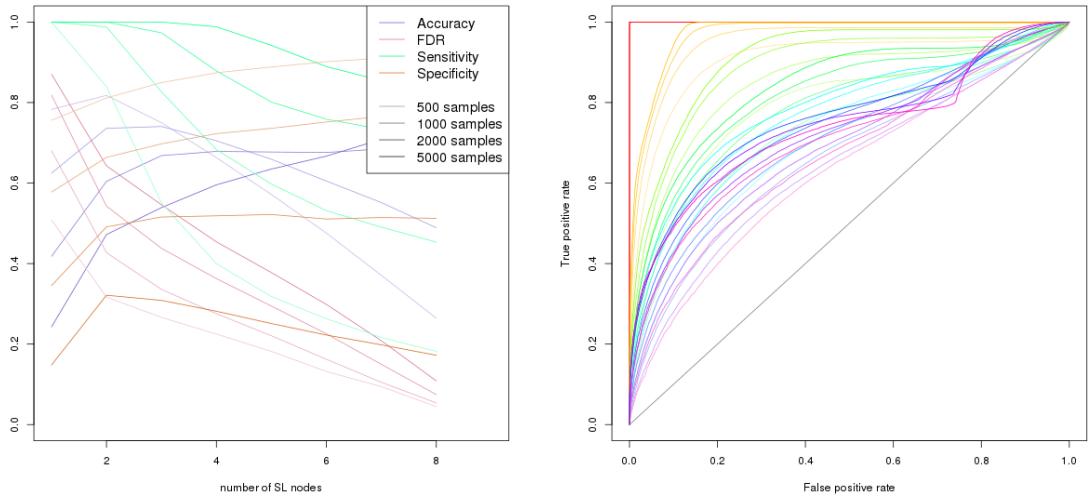


(c) Performance Between Graph Structures

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.

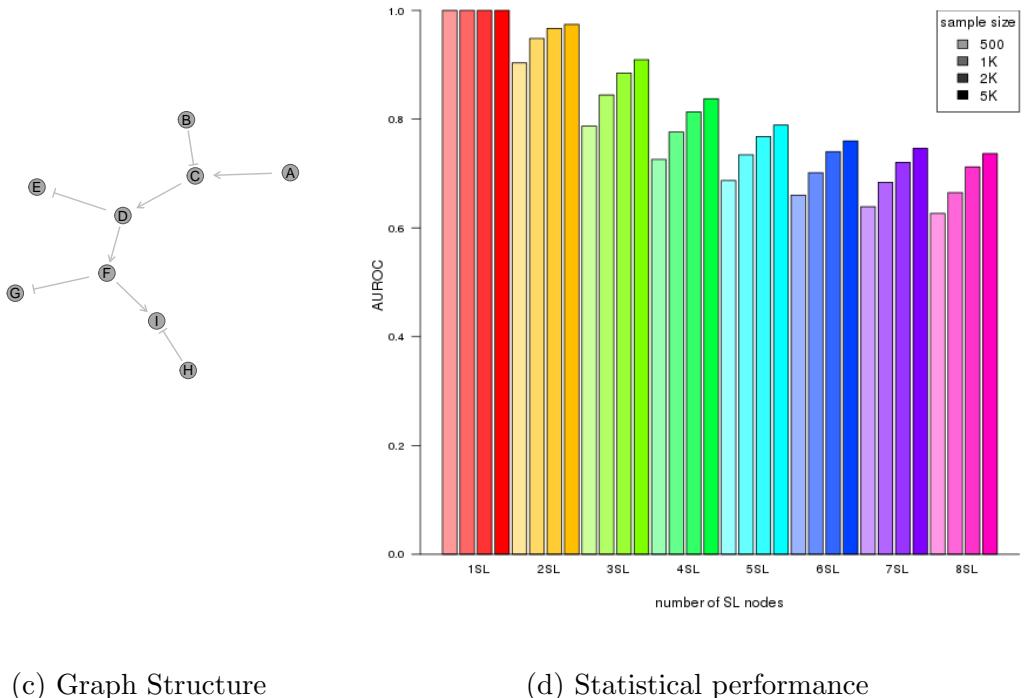
lethal genes from the correlation structure were randomly selected (without regard to their relationships), with the query gene added to ensure that conditions for synthetic lethal relationships were met.

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



(a) Statistical evaluation

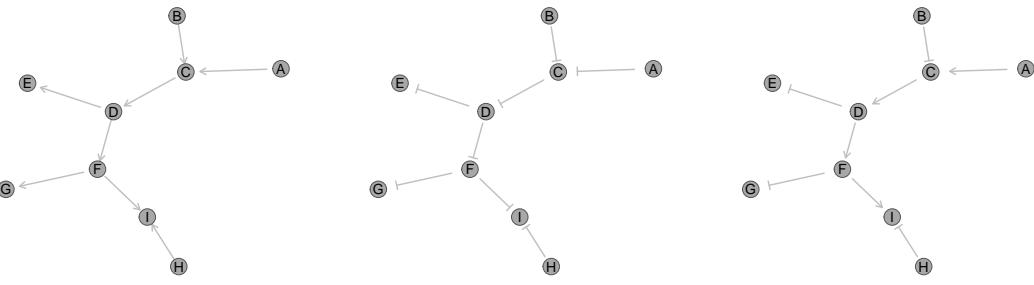
(b) Receiver operating characteristic



(c) Graph Structure

(d) Statistical performance

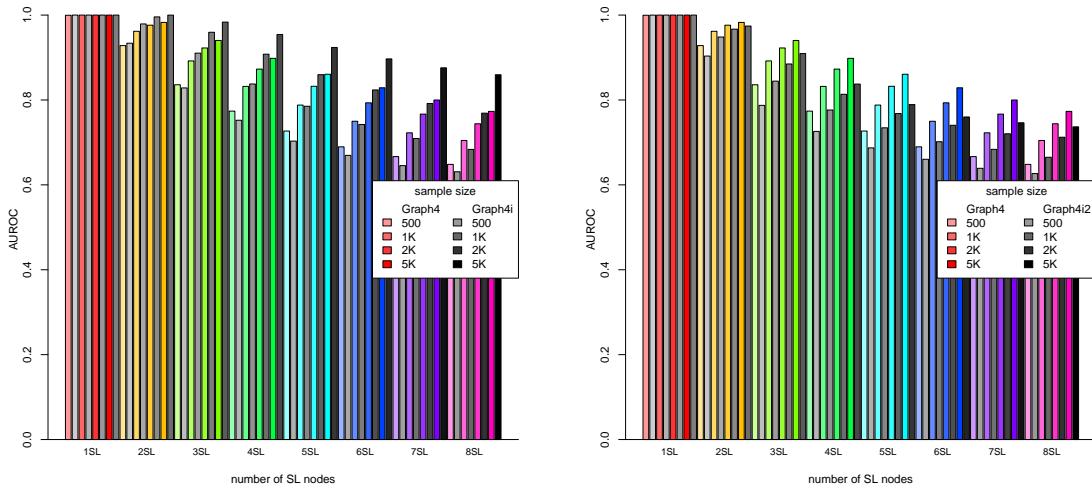
Figure 6.12: Performance of simulations on a constructed graph with inhibition.
 Simulation of synthetic lethality was performed sampling from a multivariate normal distribution generated from pathway structure with a combination of inhibitions. Performance of SLIPT declines for more synthetic partners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



(a) Activating Graph

(b) Inhibiting Graph

(c) Mixed Graph



(d) Performance Between Graphs (a) and (b) (e) Performance Between Graphs (a) and (c)

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.

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

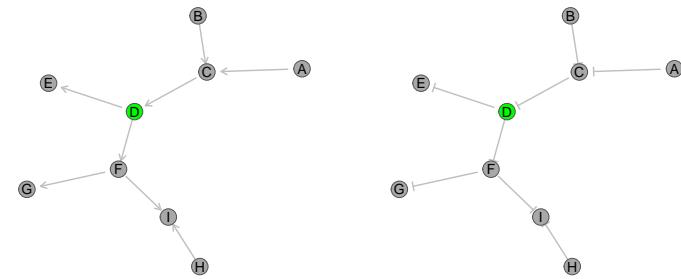
some extent) and therefore it is unlikely that they are strongly negative correlated with each other, unless they are synthetic lethal partners of each other as well. It is plausible that many synthetic lethal partner genes will serve to separately compensate for the loss of query gene function and be 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.

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



(a) Activating Graph

(b) Inhibiting Graph

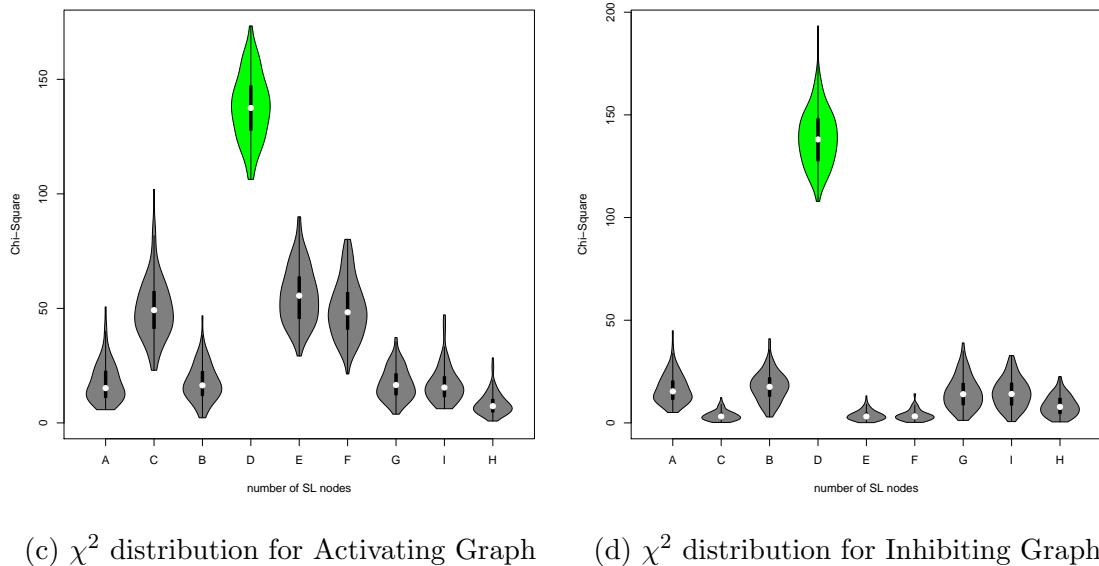


Figure 6.14: Detection of Synthetic Lethality within Graph Structure with Inhibitions. The gene “D” was designated to be synthetic lethal and the χ^2 value from SLIPT was computed for each gene across each graph structure. The χ^2 values were computed in 100 simulations of datasets of 20,000 genes including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

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 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 in-

hibitory 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 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

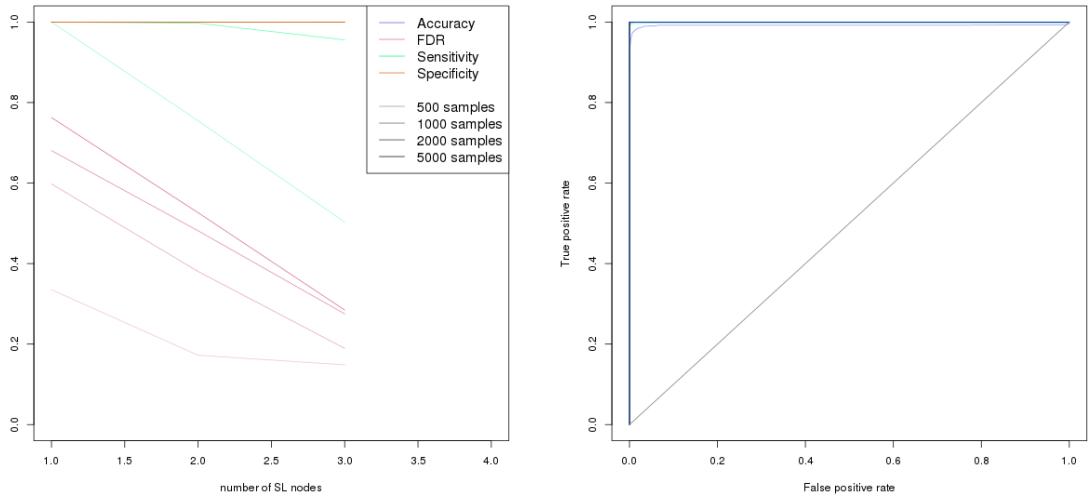
As noted in Section 6.2.1.1, the high proportion of synthetic lethal partners in small networks made accurately assessing the performance of SLIPT with higher numbers of true partners to detect (and fewer true negatives) difficult. Similarly, adding more true negative genes in previous simulations increased the performance of SLIPT, particularly the specificity to reduce the number of false positives (as shown in Sections 3.3 and 6.1). Building on these findings, here 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 serves to test the performance of SLIPT at 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 performed well across parameter values. In this case, a reduction in sensitivity was the cause of poorer performance as a higher number of non synthetic lethal genes were detected as true negative with a low false positive rate and high accuracy. This further supports the use of stringent χ^2 p-value thresholds (adjusted by FDR) for testing for synthetic lethality in gene expression data across the number of genes in human and cancer data.

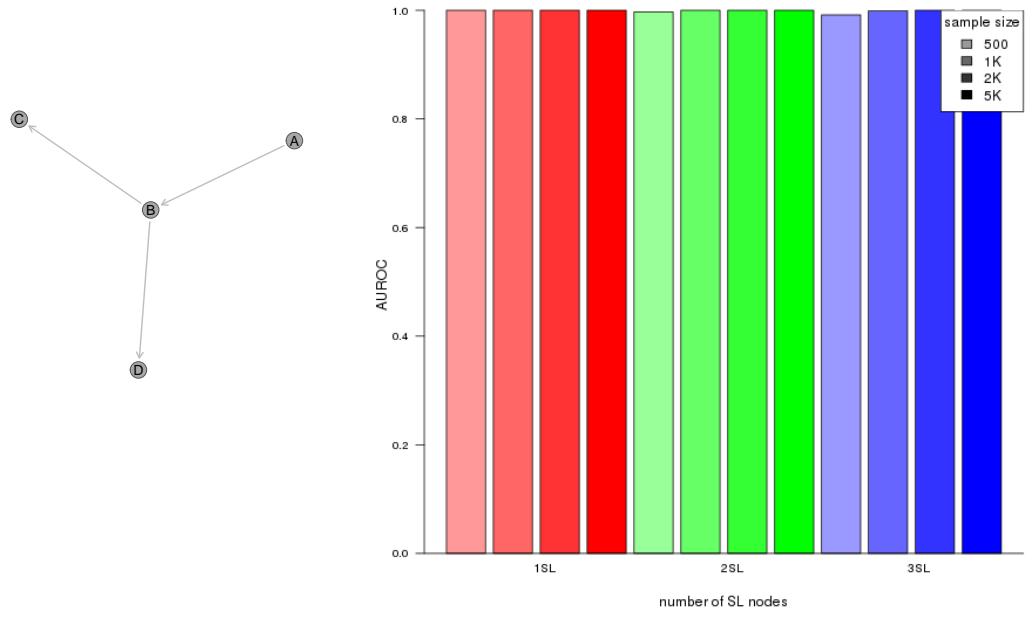
In a direct comparison with simulations in the graph structure alone (as performed in Section 6.2.1.1), detection of synthetic lethality with SLIPT performs consistently better in a larger dataset with many true negative genes to detect (as shown in Figure 6.16). This is a desirable property of the SLIPT methodology as it has a high specificity and low false positive rate. 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 negative genes to detect does not necessarily apply in an inhibiting graph structure. 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 (as shown in Figure ??). There is cause for concern since the biological pathways commonly contain inhibiting relationships (and inverse correlations), however, they are rarely as frequent as modelled here. It is reassuring that the performance in the inhibiting graph structure is comparable to simulations of the graph structure in isolation rather than diminished. It is expected that the findings based on simulations of genes with pathway structures in smaller datasets (as described in Section 6.2.1) will be relevant to larger datasets since the simulation results in these perform comparably or higher with more non synthetic lethal genes to distinguish from them even with inhibitory relationships within the graph structure



(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.

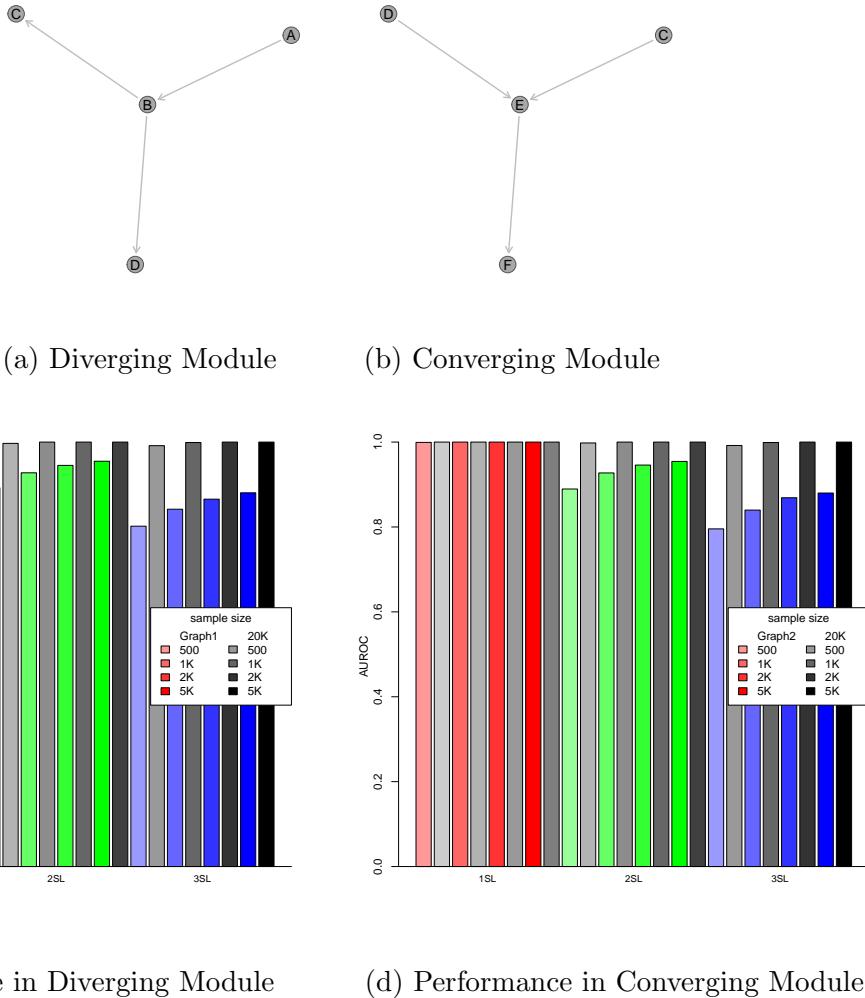


Figure 6.16: **Performance on a simple graph improves with more genes.** Simulations were performed with each of the graph structures to detect synthetic lethal partners within them. In either structure, performance of detection in a dataset containing on the graph structure (in colour) was lower than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

Performance of synthetic lethal detection of SLIPT in graphs structures with inhibitions included in a larger dataset of non synthetic lethal genes did not necessarily 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 activating graph structures within a similar dataset.

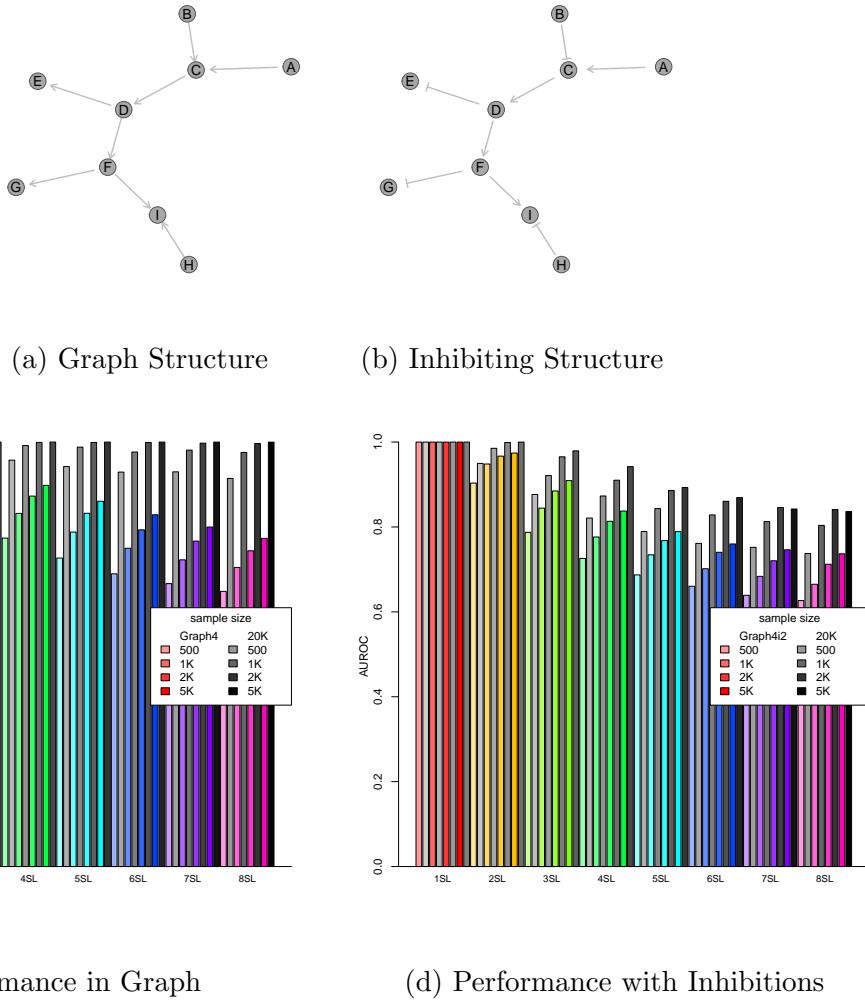


Figure 6.17: **Performance on an inhibiting graph improves with more genes.** Simulations were performed in a graph structure with activating and inhibiting relationships to detect synthetic lethal partners within them. In contrast to an activating graph, performance of detection in a dataset containing 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.

This poorer performance is unlikely to occur due to highly negatively correlated genes being false positives as they 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, 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 which may be biologically implausible (e.g., strong inhibitory relationships between them). Therefore the poorer performing inhibiting graphs within larger datasets are not cause for concern as the cases where SLIPT performs poorly are combinations of simulated synthetic lethal genes which are unlikely to occur within biological pathways. Furthermore the simulation procedure has used included higher-order synthetic lethal to produce the weakest signal of synthetic lethality for individual partner genes and these are still detectable by SLIPT.

6.3 Simulations in More Complex Graph Structures

As shown in Figure K.8, sensitivity declines over a greater range for the number of synthetic lethal partners in a larger network with a tradeoff with specificity. However, the accuracy declines for greater numbers of synthetic lethal partners and the false positive rate peaks at intermediate values. In this range, difference between simulations varies with greater sample size. The AUROC results were similar for other more complex graph structures (as shown in Figures K.9 and K.10). These graphs performed similarly to each other, although they had differences from Figure K.8 in their sensitivity and specificity at an adjusted (FDR) p-value threshold. This difference may stem from different ratios of synthetic lethal and non-synthetic lethal genes to detect, since the latter graphs (in Figures K.9 and K.10) had half the total genes to that shown in Figure K.8.

However, the graph structures (of similar size) were highly distinct and yet had similar performance profiles across parameters. Therefore SLIPT is robust across pathway structures and is more affected by the number of genes to detect and the proportion of them out of those tested. As such 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 network both 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.16). 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). As previously observed (in Figures K.9 and K.10), the proportion of underlying synthetic lethal genes to detect had a greater impact on performance of detection with SLIPT than the specific structure of the genes which was replicated with inhibiting states (in Figures K.13 and K.15) and combinations with a similar proportion of negative inhibitions (in Figures 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.

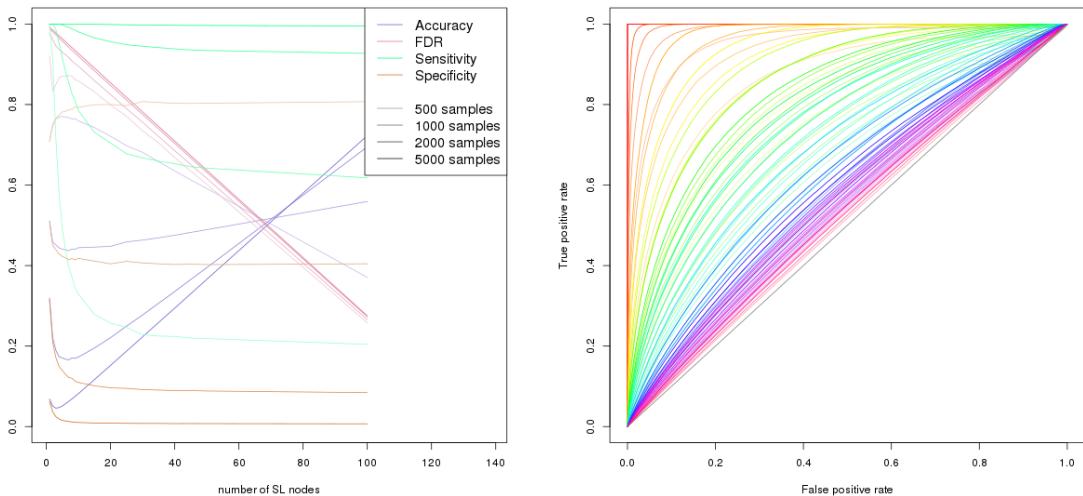
6.3.1 Simulations over Pathway-based Graphs

Thus far simulations of synthetic lethality in gene expression with correlation structures have used simple blocks of correlated genes (as used in Section 3.3) or derived from artificially constructed graph structures (as used in Section 6.2). While these are sufficient 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 extensively in Chapter 5) was used to demonstrate the simulation procedure for detecting synthetic lethality in the graph structure of a biological pathway. It is appropriate to do so since this pathway has clear directionality and signalling pathways were among those identified to be synthetic lethal candidates (in Chapter 4). The PI3K pathway having 138 genes is also of a moderate size and complexity compared to other biological pathways which is therefore suitable for comparison to previous graph structures of a similar scale (50–100 genes) with the complexity of a characteristic of a biological pathway.

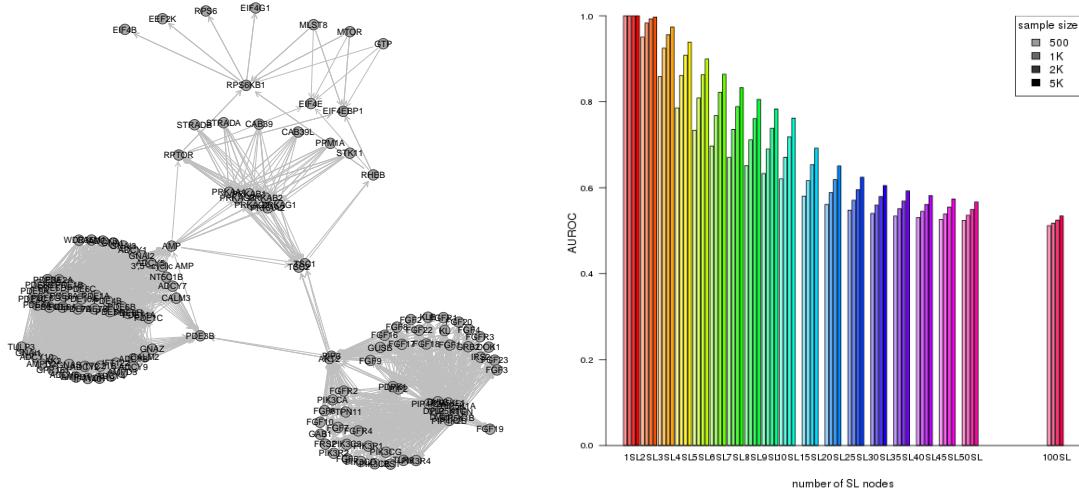
The performance of synthetic lethal detection with SLIPT in simulated expression data based on the Reactome PI3K pathway (as shown in Figure 6.18) was concordant with previous findings. SLIPT had high performance at detecting a low number of synthetic lethal genes with poorer performance for high numbers of synthetic lethal genes or lower sample sizes. In particular, the performance of simulations in the PI3K pathway was highly resembled the simulation results for constructed graphs of similar scale and complexity (as shown in Figures K.9 and K.10). 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 decreases for more synthetic lethal genes to detect within the simulated PI3K pathway, this primarily involves a reduction in sensitivity to detecting underlying synthetic lethal genes rather than false positives as the false positive rate decreases, the accuracy increases, and the specificity is relatively unperturbed (being more dependent on sample size). Thus SLIPT is stringent in biological graph structures and appropriate for detection of synthetic lethal genes in complex correlation structures in gene expression data involving biological pathways.

These simulations were replicated in the larger and more complex , 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 performed well 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.

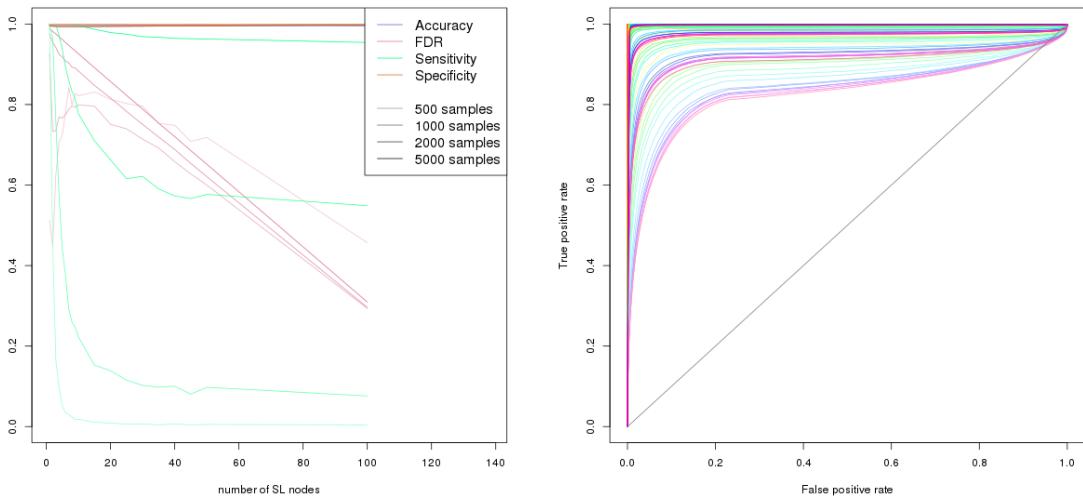
6.3.2 Pathway Structures in a Simulated Human Genome

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) have a higher specificity and therefore performance of SLIPT for detecting synthetic lethal genes was higher. The simulated PI3K pathway (as shown in Figure 6.19), is no exception with high performance across parameter values, remaining 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 is a highly stringent approach suitable for application to detecting synthetic lethal genes and pathways within highly complex expression data with biological pathway structure. In particular, the poorer performing simulations were highly stringent with low false positive rates which are an important consideration given the number of non synthetic lethal genes to distinguish in a gene expression dataset. The enrichment of true synthetic lethal partners makes SLIPT valuable for triage of candidates interacting synthetic lethal partners for further validation and for pathway analysis.

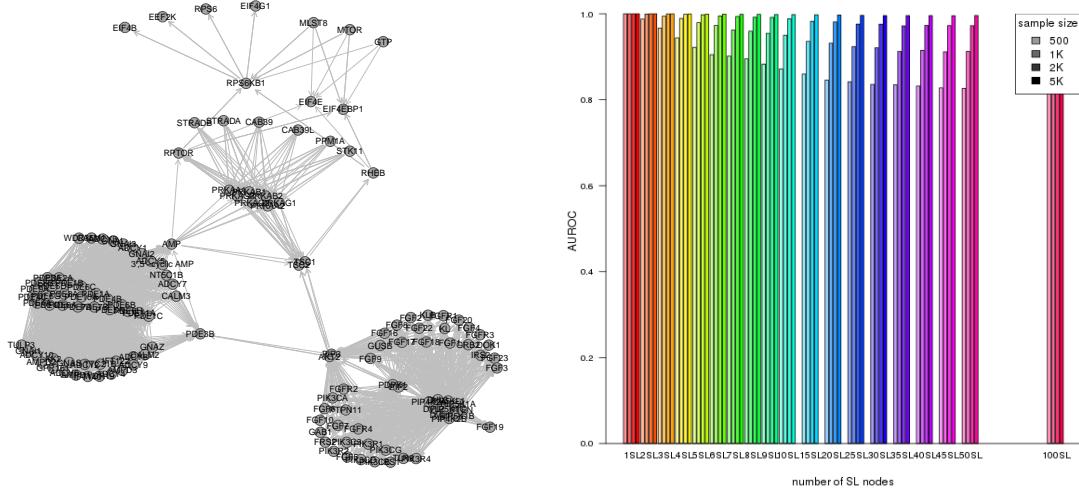
The performance of simulation of synthetic lethality within a biological pathway (e.g., the example of the PI3K cascade) was markedly higher in the context of a larger dataset of thousands of genes. As shown in a direct comparison with the graph structure alone (in Figure 6.20c), performance was consistently higher across parameters in pathways of biological complexity from the Reactome database (Croft *et al.*, 2014) such as PI3K cascade). These findings were also replicated in the larger G_{αi} signalling pathway (shown in Figures K.18 and 6.20d).

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



(a) Statistical evaluation

(b) Receiver operating characteristic

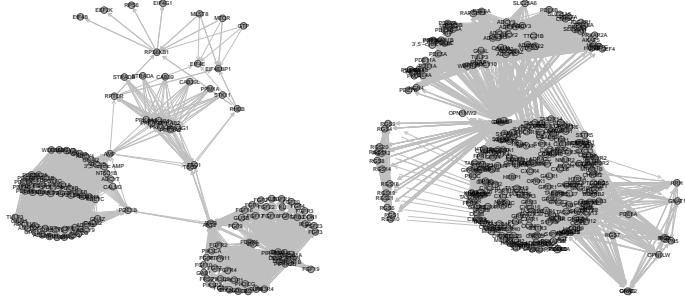


(c) Graph Structure

(d) Statistical performance

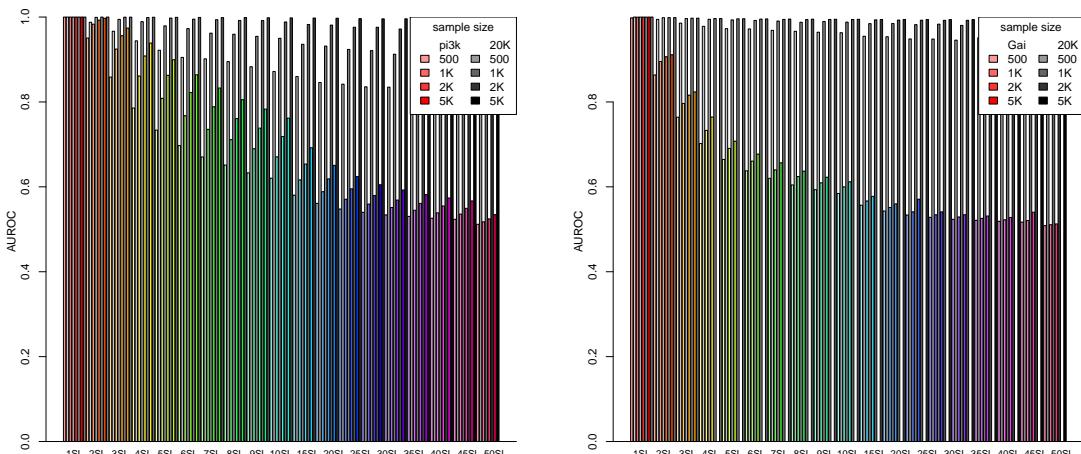
Figure 6.19: Performance of simulations including the PI3K cascade. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (without correlation structure apart from the Reactome PI3K cascade). Performance of SLIPT was high across parameters for detecting synthetic 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.

Negative genes (non synthetic lethal) inversely correlated with the underlying synthetic lethal partners will be distinguishable by SLIPT with high specificity. Since syn-



(a) The PI3K cascade

(b) G signalling



(c) Performance in the PI3K cascade

(d) Performance in G signalling

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.

thetic lethal genes are detectable will reasonable performance in large scale simulated gene expression data and highly (positively) correlated genes in pathway structures, these findings serve as a conservative lower estimate for SLIPT detecting synthetic lethal genes within a synthetic lethal biological pathway in empirical data. While synthetic lethal genes are distinguishable from correlated genes to varying extents in simulations, false positives are also more likely to be 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 statistically modelling also enables 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) using 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. A 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 enables 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 suffi-

cient 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.

Chapter 7

Discussion

This thesis combines analysis of gene expression data from TCGA with experimental screening results (Telford *et al.*, 2015) to demonstrate synthetic lethal discovery for *CDH1* in expression data generated by genomics technologies with comparisons to existing experimental candidates. Together these findings further elucidate the functions for *CDH1* in the cell, functional redundancy in breast cancer, and potential targets against cancers with loss of *CDH1* function. These candidate synthetic lethal genes were further investigated for relationships within synthetic lethal pathways, developing a network-based approach to comparing genes identified in genomics experiments and analyses in the process.

The synthetic lethal detection methodology, SLIPT, that was applied to gene expression data throughout this thesis was evaluated with simulated data. A simulation procedure was developed to stringently generate gene expression data from known synthetic lethal partners in simulated data, including simple and complex correlation structures and modelling synthetic lethal genes within pathways. Together, these results demonstrate SLIPT as a robust widely applicable gene expression analysis procedure (for which an R package has been released) for discovery of synthetic lethal partner genes. Performance of SLIPT on simulated data also highlights the strengths of the procedure and future directions to improve upon it.

7.1 Synthetic Lethality and *CDH1* Biology

The *CDH1* gene was selected to identify synthetic lethal partners to demonstrate the novel SLIPT methodology as an important tumour suppressor gene in cancers. These include sporadic breast and stomach cancers and the familial syndromes such as HDGC. The analysis of synthetic lethal partners of *CDH1* in breast and stomach cancers was

also enabled by the availability of molecular data (Bass *et al.*, 2014; TCGA, 2012) and a synthetic lethal screen conducted in MCF10A breast cells (Chen *et al.*, 2014; Telford *et al.*, 2015).

Synthetic lethal interactions are generally regarded to arise due to functional redundancy (Boone *et al.*, 2007; Fece de la Cruz *et al.*, 2015; Kaelin, Jr, 2005) and as such the synthetic lethal partners of *CDH1* indicates the wide-ranging biological functions that E-cadherin is involved in. The diverse synthetic lethal pathways identified supports the known pleiotropic nature of the *CDH1* gene by detecting established functions of *CDH1*, replicating candidates from an experimental screen (Telford *et al.*, 2015), and identifying novel interactions with candidate genes and pathways for further investigation. The highly pleiotropic functions of E-cadherin as also consistent with *CDH1* being a tumour suppressor gene for which epithelial cells are significantly disrupted at the molecular level and prone to becoming cancerous.

7.1.1 Established Functions of *CDH1*

The *CDH1* has established functions in cell-cell communication and maintaining the cytoskeleton, specifically with cell-cell adhesion by forming tight junctions and the adherens complex. More recently, additional functions of *CDH1* in the extracellular matrix and fibrin clotting have also been identified. Synthetic lethal interactions within biological pathways (i.e., partners in the same pathway as the query gene) are expected according to previous synthetic lethal experiments and (Boone *et al.*, 2007; Kelley and Ideker, 2005). Synthetic lethal interactions identified in these pathways are consistent with these being functions of *CDH1*, in addition to potentially actionable targets against cancers.

7.1.2 The Molecular Role of *CDH1* in Cancer

The involvement of *CDH1* in the extracellular matrix is also important in cancers as it indicates a mechanism by which *CDH1* loss may affect the tumour microenvironment, contributing to its role as a tumour and invasion suppressor. Furthermore, perturbations in the extracellular matrix and tumour microenvironment present an potential means by which to specifically inhibit (cancerous) *CDH1*-deficient cells in addition to those currently being considered. Few genes in extracellular pathways were detected in an experimental screen (Telford *et al.*, 2015) conducted in an isolated cell model (Chen *et al.*, 2014) but these are not expected to be detected in such as system. These may be further supported in further investigations with 3D cell culture, “organoid”, or mouse xenograft cancer models.

In contrast, many of the pathways involved in cell signalling, including GPCRs, were identified by SLIPT in addition to the experimental screen (Telford *et al.*, 2015). These support the previous results in cell line models, that these pathways are essential to growth of *CDH1*-deficient cancers and present a potential vulnerability specific to these (cancerous) cells. Furthermore, the replication of synthetic lethality of *CDH1* with cell signalling pathways in TCGA data across cancer types and genetic backgrounds robustly supports these pathways being clinically applicable beyond the genetic background of the model system of *CDH1*^{-/-} MCF10A cells (Chen *et al.*, 2014). While the specific synthetic lethal genes were not as consistently detected between the SLIPT analyses and siRNA screen (Telford *et al.*, 2015), the was sufficient to identify synthetic lethal pathways for further experimental investigation which are more likely to be replicated between genetic backgrounds (Dixon *et al.*, 2008). Together these results demonstrate how SLIPT can be integrated with an experimental screen to triage potential therapeutic targets for further pre-clinical investigation.

The analysis of expression data with SLIPT is also indicative of additional biological mechanisms of synthetic lethal in pathways beyond those identified in screening experiments (Telford *et al.*, 2015). In particular, translation and regulatory pathways, involving 3' UTRs and NMD, were identified as candidate synthetic lethal pathways with *CDH1* by SLIPT. These present downstream target regulated by the putative synthetic lethal signalling pathways which cancer cells are dependent on for sustained protein expression (Gao and Roux, 2015) to proliferate and evade host defense processes such as apoptosis and immune responses.

7.2 Significance

7.2.1 Synthetic Lethality in the Genomic Era

Development of an effective synthetic lethal discovery tool for bioinformatics analysis has a wide range of applications in genetics research including functional genomics, medical and agricultural applications. The SLIPT approach demonstrated in this thesis is widely applicable to other genes and biological questions. In addition to further query of cancer genes, including other tissues, synthetic lethal gene functions are also of wider interest for their implications for genetic redundancy. Highly redundant genes and the genetically robust systems they give rise to are of further relevance to evolutionary, developmental, and systems biology to understand how these change over time and play a role in fundamental development of cell types, in addition to cancers.

Developmental genes in particular, are highly evolutionary conserved and subject to high rates of redundancy. These are often difficult to study with conventional functional genetics since individual knockouts of redundant genes do not necessarily have a mutant phenotype. Identifying genes with a common function is therefore also important to the study of developmental genes with unknown functions. Synthetic lethal discovery methods such as SLIPT provide a genomic approach to further systematic characterisation of gene function including such highly redundant developmental genes.

Similarly, variants of unknown significance and modifier loci are a major concern in human genetics, including “monogenic” and “rare” diseases. Many of these could potentially be difficult to characterise individually due to synthetic lethal interactions where additional loci contribute to the disease (or only compensate for some variants). As such systematic identification of synthetic lethal interactions also has applications in the study of such “oligogenic” diseases along with similar applications in the study of heritability for traits including agricultural genome-based selection.

Genetic redundancy is also a concern in pharmacology. Polypharmacology and network medicine are rationales to account for this by using drugs with multiple (known and specific) targets (Barabási *et al.*, 2011; Hopkins, 2008). Further characterisation of synthetic lethal genes will be valuable to the design of effective multi-target drugs or combination therapies in a range of therapeutic applications including molecular targeted therapies against cancer for which combination therapies are a popular solution for acquired resistance against individual targeted therapies. Characterisation of genetic interactions and combination therapies also has the potential to expand pharmacogenomics investigations to understanding the impact of genotypes at multiple loci leading to adverse effects in a subset of the population or accounting for why the rest of the population does not experience this adverse effects since their synthetic lethal partner genes do not share the same variants.

Furthermore, redundant functions and synthetic lethal interactions also present a means to expand upon the concept of the “minimal” genome by accounting for essential gene functions that are performed by redundant genes (or in combination with pleiotropic) genes rather than simply those that are perturbed by individual genes as an essential gene approach is likely an underestimate that does not account for synthetic lethal interactions.

Therefore synthetic lethal interactions are a fundamentally important part of genetics and further understanding of them in a genomics context, facilitated by methods such as SLIPT, shows great potential to contribute a deeper understanding of gene

functions and their role in traits or diseases in the post-genomic era. Genes do not function in isolation and so understanding them in the context of the complexity of a cell and across genetic backgrounds (such as the data provided by TCGA) is essential to further characterise their functions and ensure that further applications are reproducible beyond experimental systems.

7.2.2 Clinical Interventions based on Synthetic Lethality

Synthetic lethal discovery with SLIPT is of particular interest in cancer research as a complementary approach to discovery of synthetic lethal drug targets. The cancer research community relies on cell line and mouse models for screening and validation experiments (Fece de la Cruz *et al.*, 2015) which would benefit from integration with gene expression analysis as demonstrated for *CDH1* and the screen conducted by Telford *et al.* (2015). The potential for synthetic lethal drug design against cancer mutations including gene loss or overexpression could lead to a revolution in cancer therapy and chemoprevention with personalised treatment of cancers and high risk individuals. Examples of the synthetic lethal strategy (Bryant *et al.*, 2005; Farmer *et al.*, 2005) for cancer treatment have been shown to be clinically effective with many large-scale RNAi screens recently conducted to aim discover gene function and drug targets for similar application with other cancer genes, including cancers in other tissues.

While SLIPT analysis and RNAi screens represent a significant step towards anti-cancer medicines, further validation is required to ensure that the synthetic lethal candidate genes and pathways identified for *CDH1* in breast and stomach cancer are applicable against *CDH1*-deficient cancers in the clinic. Validation with RNAi or pharmacological inhibitors is needed since both the SLIPT analysis and siRNA screen are susceptible to false positives. These candidates will need to be tested in pre-clinical models (cell lines and mouse xenografts) before proceeding to clinical trials. A therapeutic intervention will also require a targeted therapeutic against the synthetic lethal partner if one has not been developed against another disease (for which it can be re-purposed). Drug targets must be feasible to have effective anti-cancer interventions designed against them, which raises the need for targets with existing drugs in the clinic, trials, or feasible to development with structural analysis or screening. Druggable targets could be selected by gene functions known to be amenable to drugs, with a structure amenable with development, with conserved specific sites without homology to other genes, or with known approval or developing drugs which could be repurposed from other disease applications.

Targeted therapeutics designed based on synthetic lethal interactions have potential to vastly expand the applications of “precision medicine” against molecular targets, particularly in cancer where many have been cancer genes have been identified. Synthetic lethality expands the range of cancer genes which can be (indirectly) targeted to include tumour suppressor genes with loss of function (such as *CDH1*) and oncogenes with disrupted functions that are dysregulated or highly homologous to non-cancerous proto-oncogenes (such as *MYC*, *EGFR* or *KRAS*). Applications against tumour suppressor genes is a particularly important application as these cannot be approached by careful dosing. Synthetic lethal drug design also has the added benefit of being highly specific against a particular genotype (such as *CDH1*^{-/-}) with the potential for target therapies with a wide therapeutic index and few adverse effects, in contrast to many current anti-cancer drug regimens (Hopkins, 2008; Kaelin, Jr, 2009). These properties are highly desirable for chemoprevention applications such as treatment against *CDH1*-deficient early cancers in HDGC patients before they are detectable during screening.

7.3 Evaluating the Synthetic Lethality Prediction Tool

7.3.1 Strength of the Synthetic Lethality Prediction Tool

7.3.2 Limitations of the Synthetic Lethality Prediction Tool

7.3.3 Comparisons to Alternative Methods

7.3.3.1 Combined with Experimental Screening

7.3.3.2 Differences to Computational Methods

7.4 Future Directions

Such a bioinformatically-informed synthetic lethal screening and validation strategy could be integrated into existing and future screens for synthetic lethality in cancer.

Possible improvements to the SLIPT method include developing a Bayesian inference method or simulations and modelling to account for pathway structure among synthetic lethal genes. Another extension would be to test for higher order synthetic lethal interactions, where 3 or more genes perform a redundant function.

The synthetic lethal discovery strategy could be adapted to any form of gene inactivation or disruption such as such as changes to gene expression, regulation, epigenetics,

DNA sequence, or copy number which could plausibly induce cell death due to SL interactions. Further applications of synthetic lethal interactions such as analysis of gene networks, tissue specificity, evolutionary conservation, or drug target feasibility are possible with synthetic lethal candidates predicted with confidence on a large scale.

Further development of the synthetic lethal model and simulation is needed to explore the parameters, ensure relevance to empirical data analysis, and understanding the implications of findings so far. An example of more complex correlation structure is shown in supplementary Figures S1 and S2 with genes correlated to the Query genes (showing need for directional synthetic lethal condition) and correlated with other non-synthetic lethal genes (showing the predictions are robust to other correlation structure). The impact of these modifications on model performance in a large number of genes or simulation replicates is yet to be seen or whether such correlation structure reflects the correlation structure of empirical data (as shown in Figure 3 with the row dendrogram for correlation distance between genes), known biological pathways, or known synthetic lethal interactions. Correlation between synthetic lethal genes could also be considered.

Comparing the findings of modelling and simulation with public gene expression analysis and experimental screen targets is still needed to identify putative synthetic lethal interactions. This application will be tested with the example of CDH1 as a query gene in breast cancer for follow up to earlier results, relevance to ongoing research in the Cancer genetics Laboratory, and comparison to the experimental screen data of MCF10A cells by Telford et al. (2015). While this methodology is intended to be widely applicable, particularly to other cancer genes and will be made available to the research community (manuscript and code release in preparation).

There are several avenues for further research on synthetic lethality in breast cancer. The main alternative themes are network analysis with a focus on tissue specificity or drug feasibility with an emphasis on pharmacogenomics, biological pathways, and whether candidate targets could be inactivated by compounds with favourable pharmacokinetic properties. Either approach remains within the scope of the project, although each will require adoption of new computational tools, which is important topic for consideration in the meeting and changes to the project direction later in the year.

7.4.1 Refinements Synthetic Lethality Prediction Methods

7.4.1.1 Wider Use of Synthetic Lethality Prediction

7.4.2 Validation of Synthetic Lethal Genes and Pathways

7.4.2.1 Pre-clinical and Clinical Testing

7.4.3 Application to Further Genes and Pathways

Chapter 8

Conclusion

Synthetic lethal interactions are important for understanding gene function and development of targeted anti-cancer treatments. Synthetic lethal discovery with experimental screening is error prone and limited by the model systems in which it is performed. A bioinformatics tool to predict synthetic lethal interactions from genomics data would greatly benefit the cancer research community (and wider genetics research community). Several such tools exist, including one we have developed, but they have conflicting design and results are often inconsistent with experimental screen data. Therefore, modelling and simulation of synthetic lethality in gene expression data is needed to ensure the statistical validity of predictions. We have developed a model with correlation structure based on a Multivariate Normal distribution for which simulations detect synthetic lethality with high performance in simple cases and which has the potential to be developed to model complex correlation structure, biological pathways, or patterns observed in empirical gene expression data. The modelling, public data analysis, and experimental screen data approaches will be combined to further examine the example of CDH1 in breast cancer. Analysis of gene networks, tissue specificity, biological pathways, or drug targets remain options to explore tool development and implications for synthetic lethal cancer research in the future.

Aims

- To develop a statistical approach to detect synthetic lethal gene pairs in cancer from expression data
- To apply this methodology to public cancer gene expression data against *CDH1* and analyse pathway structure with comparisons to experimental screen data
- To construct a statistical model of synthetic lethality in multivariate normal expression data
- To develop a simulation pipeline of expression with pathway structure on a high-performance computing cluster
- To examine the statistical performance of the methodology with simulated expression including pathways and compare it to other approaches
- To release the synthetic lethal detection methodology and pathway simulation procedure as R software packages

Summary

- We have developed a Synthetic Lethal detection method that generates a high number of synthetic lethal candidates
- Pathways in cell signalling, extracellular matrix, and cytoskeletal functions were supported with experimental candidates and the known functions of E-cadherin
- Several candidate pathways were supported by mutation analysis and replicated across breast and stomach cancer
- Translation and immune functions were uniquely detected by the computational approach which may be explained by differences between patient samples and cell line models
- There remains the need to identify actionable genes within these pathways, relationships with experimental candidates, and how these pathways may affect viability when lost
- Synthetic Lethal genes were explored within a graph structures for key pathways identified previously
- In some cases these graph structures appeared to have relationships between synthetic lethal genes
- However, no existing network metrics of importance and connectivity with the networks were elevated significantly for Synthetic Lethal genes
- Nor was there significant evidence of upstream and downstream relationships between SLIPT and siRNA Candidates in a shortest path permutation analysis
- We have designed a straight-forward rational query-based synthetic lethal detection method with the example of application to *CDH1* in cancer gene expression
- We have developed a simulation pipeline to generate continuous gene expression with pathway structure including a procedure to simulate synthetic lethality

- Our simulation procedure is robust across pathway structures and has desirable performance compared to other statistical techniques

References

- Aarts, M., Bajrami, I., Herrera-Abreu, M.T., Elliott, R., Brough, R., Ashworth, A., Lord, C.J., and Turner, N.C. (2015) Functional genetic screen identifies increased sensitivity to wee1 inhibition in cells with defects in fanconi anemia and hr pathways. *Mol Cancer Ther*, **14**(4): 865–76.
- 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.
- Adamski, M.G., Gumann, P., and Baird, A.E. (2014) A method for quantitative analysis of standard and high-throughput qPCR expression data based on input sample quantity. *PLoS ONE*, **9**(8): e103917.
- Adler, D. (2005) *vioplot: Violin plot*. R package version 0.2.
- Agarwal, S., Deane, C.M., Porter, M.A., and Jones, N.S. (2010) Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol*, **6**(6): e1000817.
- Agrawal, N., Akbani, R., Aksoy, B.A., Ally, A., Arachchi, H., Asa, S.L., Auman, J.T., Balasundaram, M., Balu, S., Baylin, S.B., *et al.* (2014) Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**(3): 676–690.
- 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 Padiatrica*, **96**(5): 644–647.
- American Cancer Society (2017) Genetics and cancer. <https://www.cancer.org/cancer/cancer-causes/genetics.html>. Accessed: 22/03/2017.

American Society for Clinical Oncology (ASCO) (2017) The genetics of cancer. <http://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer>. 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.

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.].
- Beck, T.F., Mullikin, J.C., and Biesecker, L.G. (2016) Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. *Clin Chem*, **62**(4): 647–654.

- 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.
- Boettcher, M., Lawson, A., Ladenburger, V., Fredebohm, J., Wolf, J., Hoheisel, J.D., Frezza, C., and Shlomi, T. (2014) High throughput synthetic lethality screen reveals a tumorigenic role of adenylate cyclase in fumarate hydratase-deficient cancer cells. *BMC Genomics*, **15**: 158.
- 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.
- 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.
- 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.
- Burk, R.D., Chen, Z., Saller, C., Tarvin, K., Carvalho, A.L., Scapulatempo-Neto, C., Silveira, H.C., Fregnani, J.H., Creighton, C.J., Anderson, M.L., *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**(7645): 378–384.
- 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.
- Cancer Research UK (2017) Family history and cancer genes. <http://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/family-history-and-inherited-cancer-genes>. Accessed: 22/03/2017.
- 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, K., Yang, D., Li, X., Sun, B., Song, F., Cao, W., Brat, D.J., Gao, Z., Li, H., Liang, H., *et al.* (2015) Mutational landscape of gastric adenocarcinoma in Chinese: implications for prognosis and therapy. *Proc Natl Acad Sci USA*, **112**(4): 1107–1112.

Chen, S. and Parmigiani, G. (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol*, **25**(11): 1329–1333.

Chen, X. and Tompa, M. (2010) Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotechnol*, **28**(6): 567–572.

Cherniack, A.D., Shen, H., Walter, V., Stewart, C., Murray, B.A., Bowlby, R., Hu, X., Ling, S., Soslow, R.A., Broaddus, R.R., *et al.* (2017) Integrated Molecular Characterization of Uterine Carcinosarcoma. *Cancer Cell*, **31**(3): 411–423.

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 -Opioid Agonist-Mediated Adenylyl Cyclase Supersensitization. *Journal of Pharmacology and Experimental Therapeutics*, **310**(1): 215–222.

Clough, E. and Barrett, T. (2016) The Gene Expression Omnibus Database. *Methods Mol Biol*, **1418**: 93–110.

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.

- Collins, F.S., Morgan, M., and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, **300**(5617): 286–290.
- 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.
- Corcoran, R.B., Ebi, H., Turke, A.B., Coffee, E.M., Nishino, M., Cogdill, A.P., Brown, R.D., Della Pelle, P., Dias-Santagata, D., Hung, K.E., *et al.* (2012) Egfr-mediated reactivation of mapk signaling contributes to insensitivity of *BRAF*-mutant colorectal cancers to raf inhibition with vemurafenib. *Cancer Discovery*, **2**(3): 227–235.
- 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.
- Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403): 346–352.

- 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.
- 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.
- Dienstmann, R. and Tabernero, J. (2011) *BRAF* as a target for cancer therapy. *Anti-cancer Agents Med Chem*, **11**(3): 285–95.
- 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.
- Dorogovtsev, S.N. and Mendes, J.F. (2003) *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, USA.

- 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.
- Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014) Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, **23**(22): 5866.
- 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.
- 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 <https://CRAN.R-project.org/package=mvtnorm>.
- Gilbert, W. and Maxam, A. (1973) The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences*, **70**(12): 3581–3584.
- Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P., and Caldas, C. (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, **16**(5): 991–1006.
- Globus (Globus) (2017) Research data management simplified. <https://www.globus.org/>. Accessed: 25/03/2017.
- 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.
- 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., Scouler, 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 News*, **11**(1): 10–18.
- Hamerman, 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.
- Han, J.D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995): 88–93.
- 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.
- Heiskanen, M., Bian, X., Swan, D., and Basu, A. (2014) caArray microarray database in the cancer biomedical informatics grid™ (caBIG™). *Cancer Research*, **67**(9 Supplement): 3712–3712.
- 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. problémes combinatorics at theorie des graphes. *Proc Coll Int CNRS, Orsay*, **260**: 219–223.
- Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., Chandrasekharan, S., *et al.* (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol*, **8**(5): R76.
- 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.
- Holme, P. and Kim, B.J. (2002) Growing scale-free networks with tunable clustering. *Physical Review E*, **65**(2): 026107.
- 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.
- Illumina, Inc (Illumina) (2017) Sequencing and array-based solutions for genetic research. <https://www.illumina.com/>. Accessed: 26/03/2017.

- International HapMap 3 Consortium (HapMap) (2003) The International HapMap Project. *Nature*, **426**(6968): 789–796.
- Internationl Human Genome Sequencing Consortium (IHGSC) (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011): 931–945.
- 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, lkopf, 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.
- Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., Yamamoto, S., Tatsuno, K., Katoh, H., Watanabe, Y., *et al.* (2014) Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. *Nat Genet*, **46**(6): 583–587.
- Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**(1): 7–15.
- Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**(7447): 67–73.
- 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., Chen, A., Guilford, P., and Black, M. (2017a) Synthetic lethal interaction prediction of target pathways in E-cadherin deficient breast cancers. Submitted to *BMC Genomics*.
- 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. (2017b) 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*.
- 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.
- 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.
- Lawrence, M.S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Gabriel, S.B., Getz, G., Ally, A., Balasundaram, M., Birol, I., et al. (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**(7536): 576–582.
- 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., Boulier, 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, 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 anticancer 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) Comprehen-

- sive 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., Hinchshaw, 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.
- 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.
- Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Science*, **74**(2): 560–564.

- McCourt, C.M., McArt, D.G., Mills, K., Catherwood, M.A., Maxwell, P., Waugh, D.J., Hamilton, P., O’Sullivan, J.M., and Salto-Tellez, M. (2013) Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. *PLoS ONE*, **8**(7): e69604.
- 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., Mastrogiannakis, 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.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7): 621–628.
- 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.
- 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.
- Peltonen, L. and McKusick, V.A. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, **291**(5507): 1224–1229.
- 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.
- Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R., et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**(7278): 191–196.
- 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.
- Prahallas, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R.L., Bardelli, A., and Bernards, R. (2012) Unresponsiveness of colon cancer to *BRAF*(v600e) inhibition through feedback activation of egfr. *Nature*, **483**(7387): 100–3.
- R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.3.2.

- Ravnan, M.C. and Matalka, M.S. (2012) Vemurafenib in patients with *BRAF* v600e mutation-positive advanced melanoma. *Clin Ther*, **34**(7): 1474–86.
- 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.
- Robin, J.D., Ludlow, A.T., LaRanger, R., Wright, W.E., and Shay, J.W. (2016) Comparison of DNA Quantification Methods for Next Generation Sequencing. *Sci Rep*, **6**: 24067.
- Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**(3): R25.
- 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.
- Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. *Nat Rev Genet*, **14**(2): 89–99.
- Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., *et al.* (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res*, **41**(Database issue): D987–990.
- Ryan, C., Lord, C., and Ashworth, A. (2014) Daisy: Picking synthetic lethals from cancer genomes. *Cancer Cell*, **26**(3): 306–308.
- Sander, J.D. and Joung, J.K. (2014) Crispr-cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, **32**(4): 347–55.
- Sanger, F. and Coulson, A. (1975) A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, **94**(3): 441 – 448.
- 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.
- Sun, C., Wang, L., Huang, S., Heynen, G.J.J.E., Prahallad, A., Robert, C., Haanen, J., Blank, C., Wesseling, J., Willems, S.M., et al. (2014) Reversible and adaptive resistance to *BRAF*(v600e) inhibition in melanoma. *Nature*, **508**(7494): 118–122.
- Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*, **27**(2): 199–204.
- 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) (2017a) The Cancer Genome Atlas Project. <https://cancergenome.nih.gov/>. Accessed: 26/03/2017.
- The Cancer Genome Atlas Research Network (TCGA) (2017b) The Cancer Genome Atlas Project Data Portal. <https://tcga-data.nci.nih.gov/>. Accessed: 06/02/2017 (via cBioPortal).
- The Cancer Society of New Zealand (Cancer Society of NZ) (2017) What is cancer? <https://otago-southland.cancernz.org.nz/en/cancer-information/other-links/what-is-cancer-3/>. Accessed: 22/03/2017.
- 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 Internation Cancer Genome Consortium (ICGC) (2017) ICGC Data Portal. <https://dcc.icgc.org/>. Accessed: 06/02/2017.
- 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.
- The Pharmaceutical Management Agency (PHARMAC) (2016) Approval of multi-product funding proposal with roche.
- 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.
- Travers, J. and Milgram, S. (1969) An experimental study of the small world problem. *Sociometry*, **32**(4): 425–443.
- Tsai, H.C., Li, H., Van Neste, L., Cai, Y., Robert, C., Rassool, F.V., Shin, J.J., Harbom, K.M., Beaty, R., Pappou, E., *et al.* (2012) Transient low doses of dna-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell*, **21**(3): 430–46.
- 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(adt-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.
- van der Meer, R., Song, H.Y., Park, S.H., Abdulkadir, S.A., and Roh, M. (2014) RNAi screen identifies a synthetic lethal interaction between PIM1 overexpression and PLK1 inhibition. *Clinical Cancer Research*, **20**(12): 3211–3221.
- van Steen, K. (2012) Travelling the world of genegene 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.
- Vargas, J.J., Gusella, G., Najfeld, V., Klotman, M., and Cara, A. (2004) Novel integrase-defective lentiviral episomal vectors for gene transfer. *Hum Gene Ther*, **15**: 361–372.
- 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.
- 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, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., *et al.* (2014) Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet*, **46**(6): 573–582.
- 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 portal a 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.

Appendix A

Sample Quality

A.1 Sample Correlation

Samples were excluded from expression analysis based on sample correlations and the clustering analysis presented below, as described in Section 2.2.2.

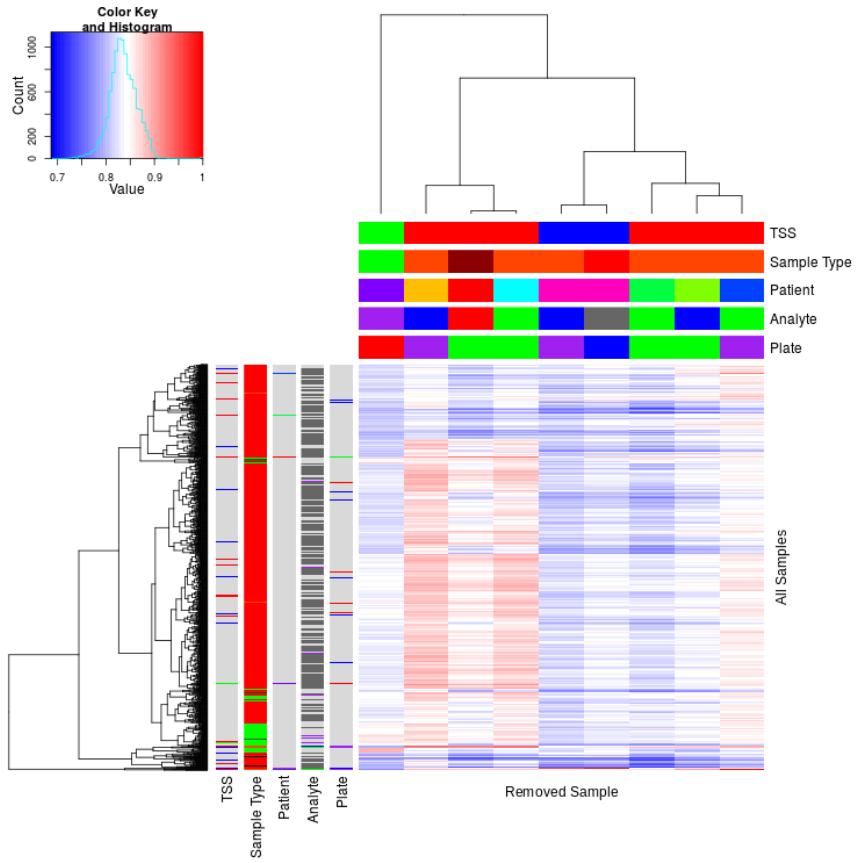


Figure A.1: Correlation profiles of removed samples. Correlation matrix heatmap (Euclidean distance) of all samples in TCGA breast cancer dataset (left) clustered for all samples against removed samples (top): tissue source site (TSS), sample type with reds for tumour and greens for normal, patient (A2QH in pink), with varied analyte and plate (corresponding to batch in Table 2.1). Excluded samples cluster at the bottom and annotation (left) show shared properties between samples in the dataset.

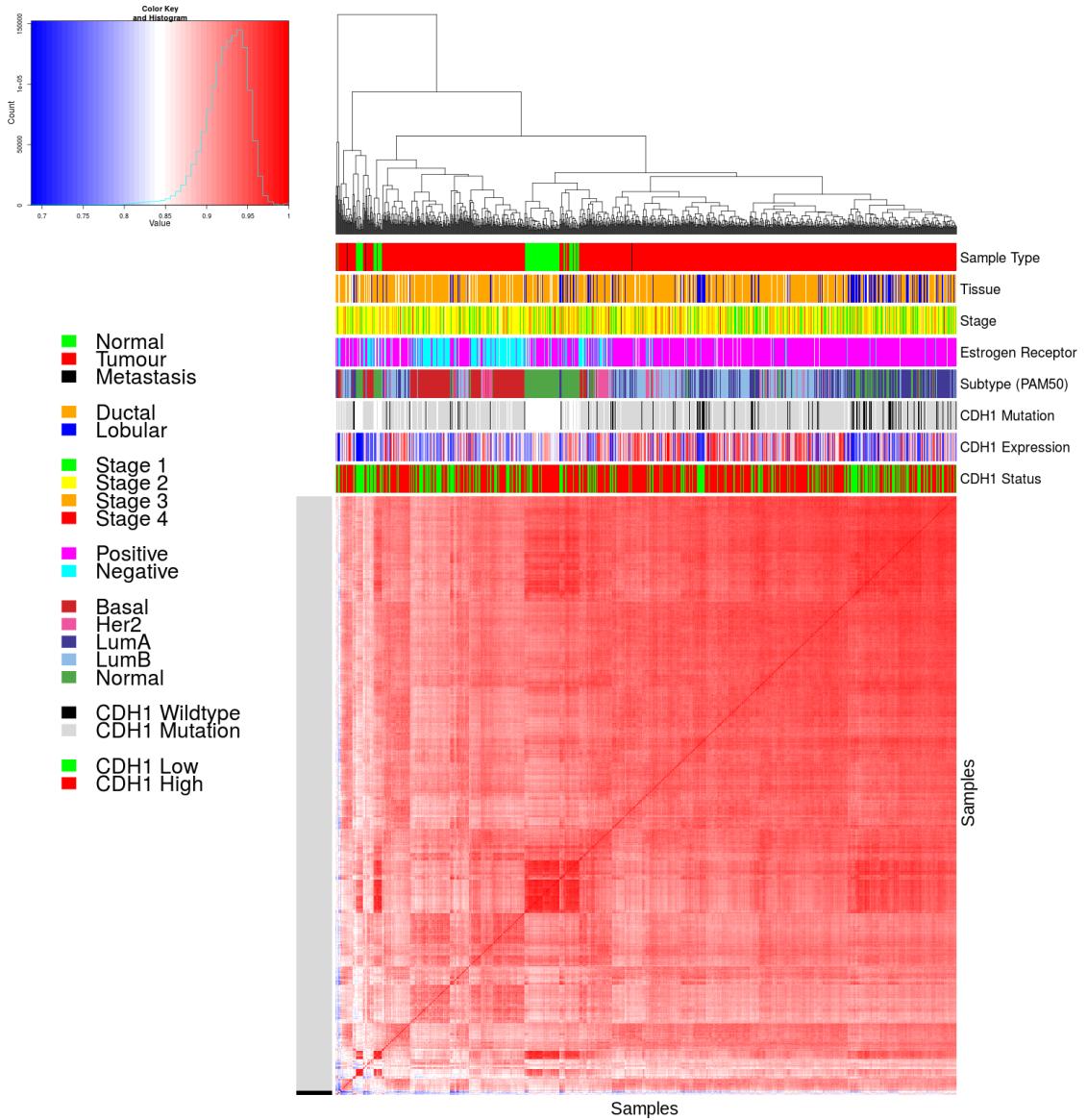


Figure A.2: Correlation analysis and sample removal. Correlation matrix heatmap (Euclidean distance) of all samples in TCGA breast cancer dataset against each other annotated for sample clinical data: sample type, tissue type, tumour stage, Estrogen receptor (IHC) and intrinsic subtype (from the PAM50 method). CDH1 somatic mutation, gene expression, and status for SLIPT prediction are also annotated. Discrete variables are coloured as displayed in the legend and continuous variables on a blue-red scale as shown in the colour key. Trimmed samples cluster at the bottom of the heatmap and the colour bars of the left show which were removed for quality concerns.

A.2 Replicate Samples in TCGA Breast

Replicate samples were picked where possible from the TCGA breast cancer gene expression data to examine for sample quality. Independent samples of the same tumour are expected to have very high Pearson's correlation between their expression profiles unless there were issues with sample collection or preparation and are thus an indicator of sample quality. The log-transformed raw read counts for replicate samples were examined in Figures A.3–A.5. These were examined before normalisation which would be expected to increase sample concordance.

Another consideration are the samples which were removed for quality concerns (in Section 2.2.2). While these were selected by unbiased hierarchical clustering (See Figure A.2), it is notable that many of the excluded (tumour) samples were performed in replicate despite relatively few replicate samples in the overall dataset. These samples correlate poorly with the rest of the dataset, in addition to with replicate samples.

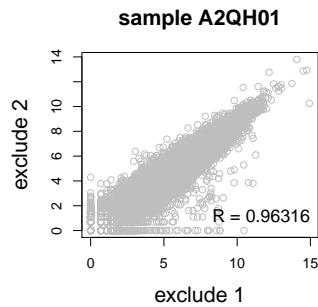


Figure A.3: **Replicate excluded samples.** Both tumour samples of patient A2QH were excluded as they were poorly correlated with other samples, although they are highly similar to each other as shown by Pearson's correlation of log-raw counts.

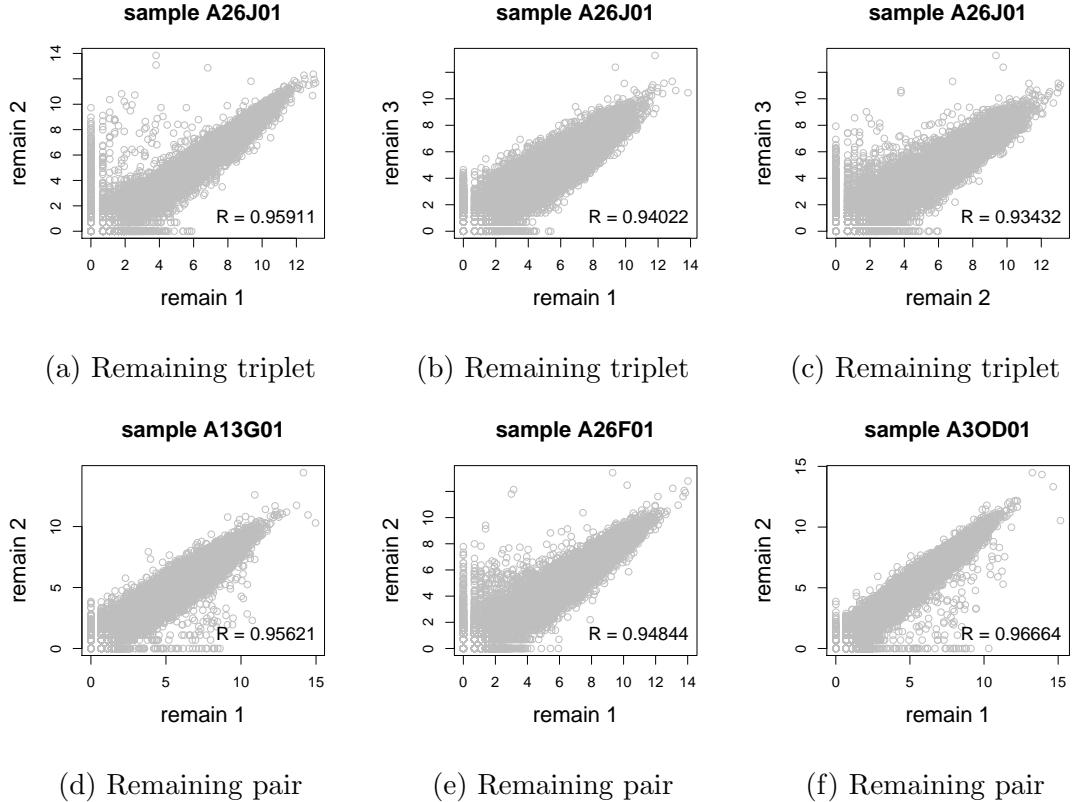


Figure A.4: **Replicate samples with all remaining.** Patient A26J was sampled 3 times and compared pairwise. Pairs of samples were also compared for other patients with replicate samples. In all cases, replicate samples remaining in the dataset were highly concordant as shown by Pearson's correlation of log-raw counts.

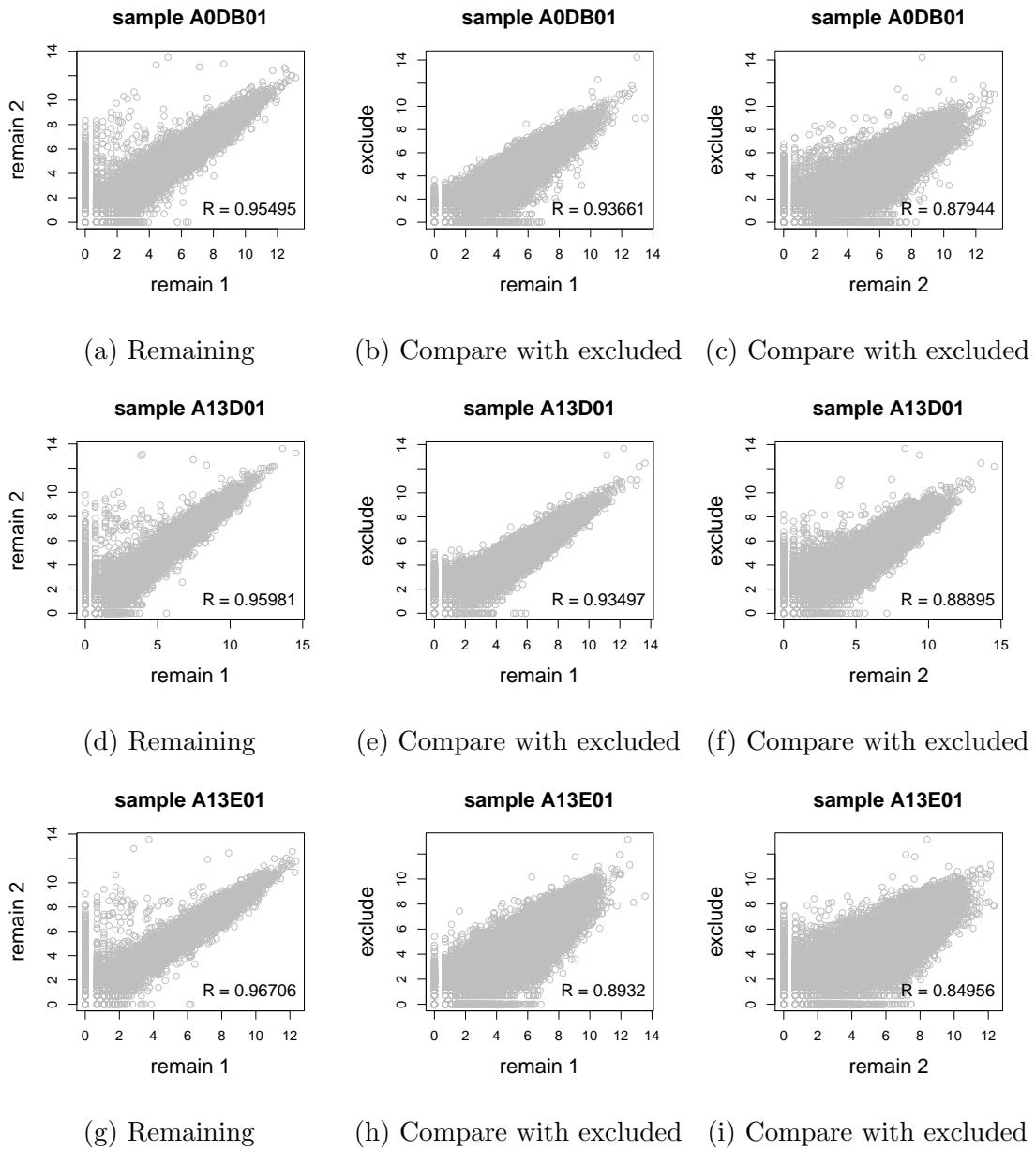


Figure A.5: **Replicate samples with some excluded.** (continued on next page)

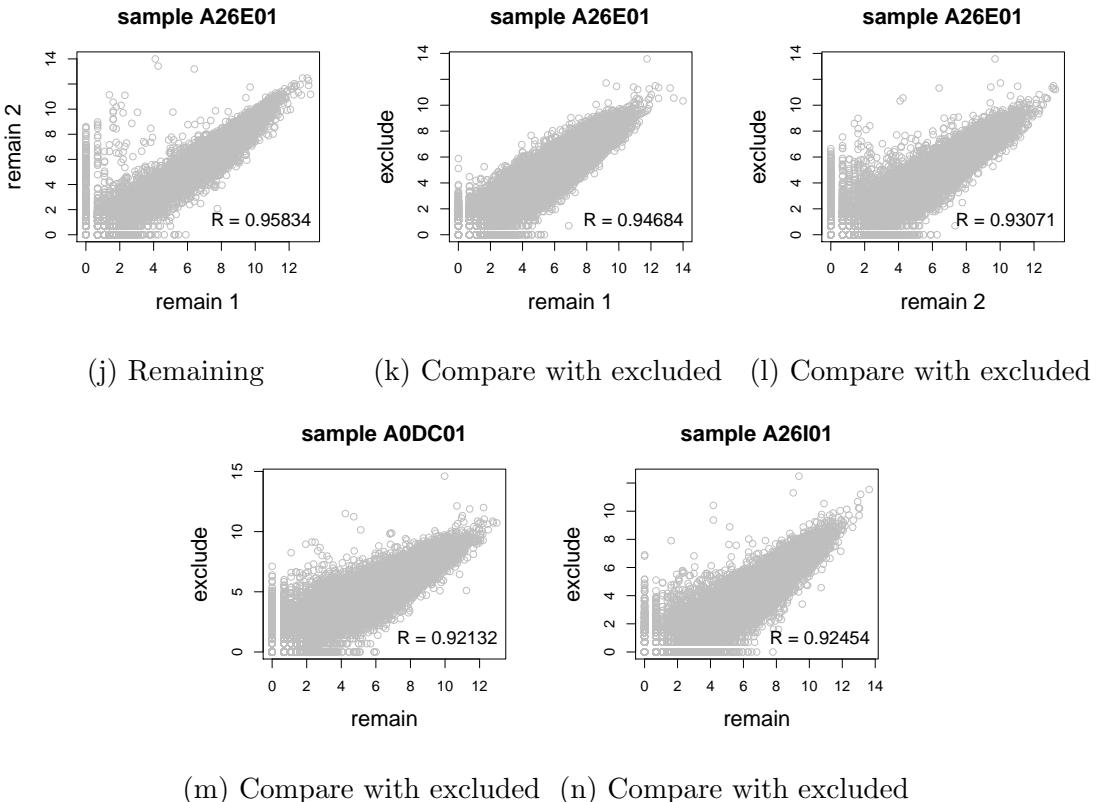


Figure A.5: **Replicate samples with some excluded.** Patients A0DB, A13D, A13E, and A26E were each sampled 3 times and compared pairwise. Pairs of samples were also compared for other patients with replicate samples. In all cases, the replicate samples remaining in the dataset more were highly concordant (as shown by Pearson's correlation of log-raw counts) than those excluded from the analysis.

Appendix B

Software Used for Thesis

Table B.1: R Packages used during Thesis

Package	Repository	Laptop	Lab	Server	NeSI
base	base	3.3.2	3.3.2	3.3.1	3.3.0
abind	CRAN		1.4-5		1.4-3
acepack	CRAN		1.4.1		1.3-3.3
ade4	CRAN		1.7-5		
annaffy	Bioconductor		1.46.0		
AnnotationDbi	Bioconductor		1.36.0	1.36.0	1.34.4
apComplex	CRAN		2.40.0		
ape	CRAN		4		3.4
arm	CRAN		1.9-3		
assertthat	CRAN	0.1	0.1	0.1	0.1
backports	CRAN	1.0.5	1.0.4	1.0.5	1.0.2
base64	CRAN			2	2
base64enc	CRAN		0.1-3		0.1-3
beanplot	CRAN		1.2	1.2	1.2
BH	CRAN	1.60.0-2	1.62.0-1	1.62.0-1	1.60.0-2
Biobase	Bioconductor		2.34.0	2.34.0	2.32.0
BiocGenerics	Bioconductor		0.20.0	0.20.0	0.18.0
BiocInstaller	Bioconductor		1.24.0	1.20.3	1.22.3
BiocParallel	Bioconductor		1.8.1	1.8.1	
Biostrings	Bioconductor		2.42.1	2.42.0	
BiSEp	Bioconductor		2.0.1	2.0.1	2.0.1

bitops	CRAN	1.0-6	1.0-6	1.0-6	1.0-6
boot	base	1.3-18	1.3-18	1.3-18	1.3-18
brew	CRAN	1.0-6	1.0-6	1.0-6	1.0-6
broom	CRAN	0.4.1			
caTools	CRAN	1.17.1	1.17.1	1.17.1	1.17.1
cgdssr	CRAN		1.2.5		
checkmate	CRAN		1.8.2		1.7.4
chron	CRAN	2.3-47	2.3-48	2.3-50	2.3-47
class	base	7.3-14	7.3-14	7.3-14	7.3-14
cluster	base	2.0.5	2.0.5	2.0.5	2.0.4
coda	CRAN		0.19-1		0.18-1
codetools	base	0.2-15	0.2-15	0.2-15	0.2-14
colorRamps	CRAN		2.3		
colorspace	CRAN	1.2-6	1.3-2	1.3-2	1.2-6
commonmark	CRAN	1.1		1.2	
compiler	base	3.3.2	3.3.2	3.3.1	3.3.0
corpcor	CRAN		1.6.8	1.6.8	1.6.8
Cprob	CRAN		1.2.4		
crayon	CRAN	1.3.2	1.3.2	1.3.2	1.3.2
crop	CRAN		0.0-2	0.0-2	
curl	CRAN	1.2	2.3	2.3	0.9.7
d3Network	CRAN		0.5.2.1		
data.table	CRAN	1.9.6	1.10.0	1.10.1	1.9.6
data.tree	CRAN		0.7.0	0.7.0	
datasets	base	3.3.2	3.3.2	3.3.1	3.3.0
DBI	CRAN	0.5-1	0.5-1	0.5-1	0.5-1
dendextend	CRAN	1.4.0	1.4.0	1.4.0	
DEoptimR	CRAN	1.0-8	1.0-8	1.0-8	1.0-4
desc	CRAN	1.1.0		1.1.0	
devtools	CRAN	1.12.0	1.12.0	1.12.0	1.12.0
DiagrammeR	CRAN		0.9.0	0.9.0	
dichromat	CRAN	2.0-0	2.0-0	2.0-0	2.0-0
digest	CRAN	0.6.10	0.6.11	0.6.12	0.6.9
diptest	CRAN	0.75-7	0.75-7	0.75-7	
doParallel	CRAN	1.0.10	1.0.10	1.0.10	1.0.10

dplyr	CRAN	0.5.0	0.5.0	0.5.0	0.5.0
ellipse	CRAN		0.3-8	0.3-8	0.3-8
evaluate	CRAN		0.1	0.1	0.9
fdrtool	CRAN		1.2.15		
fields	CRAN		8.1		
flexmix	CRAN	2.3-13	2.3-13	2.3-13	
forcats	CRAN	0.2.0			
foreach	CRAN	1.4.3	1.4.3	1.4.3	1.4.3
foreign	base	0.8-67	0.8-67	0.8-67	0.8-66
formatR	CRAN		1.4	1.4	1.4
Formula	CRAN		1.2-1		1.2-1
fpc	CRAN	2.1-10	2.1-10	2.1-10	
futile.logger	CRAN		1.4.3	1.4.3	1.4.1
futile.options	CRAN		1.0.0	1.0.0	1.0.0
gdata	CRAN	2.17.0	2.17.0	2.17.0	2.17.0
geepack	CRAN		1.2-1		
GenomeInfoDb	Bioconductor		1.10.2	1.10.1	
GenomicAlignments	Bioconductor		1.10.0	1.10.0	
GenomicRanges	Bioconductor		1.26.2	1.26.1	
ggm	CRAN		2.3		
ggplot2	CRAN	2.1.0	2.2.1	2.2.1	2.1.0
git2r	CRAN	0.15.0	0.18.0	0.16.0	0.15.0
glasso	CRAN		1.8		
GO.db	Bioconductor		3.4.0	3.2.2	3.3.0
GOSemSim	Bioconductor		2.0.3	1.28.2	1.30.3
gplots	CRAN	3.0.1	3.0.1	3.0.1	3.0.1
graph	Bioconductor		1.52.0		
graphics	base	3.3.2	3.3.2	3.3.1	3.3.0
graphsim	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
grDevices	base	3.3.2	3.3.2	3.3.1	3.3.0
grid	base	3.3.2	3.3.2	3.3.1	3.3.0
gridBase	CRAN	0.4-7	0.4-7	0.4-7	0.4-7
gridExtra	CRAN	2.2.1	2.2.1	2.2.1	2.2.1
gridGraphics	CRAN		0.1-5		

gtable	CRAN	0.2.0	0.2.0	0.2.0	0.2.0
gtools	CRAN	3.5.0	3.5.0	3.5.0	3.5.0
haven	CRAN	1.0.0			
heatmap.2x	GitHub TomKellyGenetics	0.0.0.9000	0.0.0.9000	0.0.0.9000	0.0.0.9000
hg133plus2.db	Bioconductor		3.2.3		
highr	CRAN		0.6	0.6	0.6
Hmisc	CRAN		4.0-2	4.0-2	3.17-4
hms	CRAN	0.2	0.3		
htmlTable	CRAN		1.8	1.9	
htmltools	CRAN	0.3.5	0.3.5	0.3.5	0.3.5
htmlwidgets	CRAN		0.8	0.8	
httpuv	CRAN	1.3.3		1.3.3	
httr	CRAN	1.2.1	1.2.1	1.2.1	1.1.0
huge	CRAN		1.2.7		
hunspell	CRAN		2.3		2
hypergraph	CRAN		1.46.0		
igraph	CRAN	1.0.1	1.0.1	1.0.1	1.0.1
igraph.extensions	GitHub TomKellyGenetics	0.1.0.9001	0.1.0.9001	0.1.0.9001	0.1.0.9001
influenceR	CRAN		0.1.0	0.1.0	
info.centrality	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
IRanges	Bioconductor		2.8.1	2.8.1	2.6.1
irlba	CRAN	2.1.1	2.1.2	2.1.2	2.0.0
iterators	CRAN	1.0.8	1.0.8	1.0.8	1.0.8
jpeg	CRAN		0.1-8		
jsonlite	CRAN	1.1	1.2	1.3	0.9.20
KEGG.db	Bioconductor		3.2.3		
kernlab	CRAN	0.9-25	0.9-25	0.9-25	
KernSmooth	base	2.23-15	2.23-15	2.23-15	2.23-15
knitr	CRAN		1.15.1	1.15.1	1.14
labeling	CRAN	0.3	0.3	0.3	0.3
lambda.r	CRAN		1.1.9	1.1.9	1.1.7
lattice	base	0.20-34	0.20-34	0.20-34	0.20-33

latticeExtra	CRAN		0.6-28		0.6-28
lava	CRAN		1.4.6		
lavaan	CRAN		0.5-22		
lazyeval	CRAN	0.2.0	0.2.0	0.2.0	0.2.0
les	CRAN		1.24.0		
lgtdl	CRAN		1.1.3		
limma	Bioconductor		3.30.7	3.30.3	
lme4	CRAN		1.1-12		1.1-12
lubridate	CRAN	1.6.0			
magrittr	CRAN	1.5	1.5	1.5	1.5
maps	CRAN		3.1.1		
markdown	CRAN		0.7.7	0.7.7	0.7.7
MASS	base	7.3-45	7.3-45	7.3-45	7.3-45
Matrix	base	1.2-7.1	1.2-7.1	1.2-8	1.2-6
matrixcalc	CRAN	1.0-3	1.0-3	1.0-3	1.0-3
mclust	CRAN	5.2	5.2.1	5.2.2	5.2
memoise	CRAN	1.0.0	1.0.0	1.0.0	1.0.0
methods	base	3.3.2	3.3.2	3.3.1	3.3.0
mgcv	base	1.8-16	1.8-16	1.8-17	1.8-12
mi	CRAN		1		
mime	CRAN	0.5	0.5	0.5	0.4
minqa	CRAN		1.2.4		1.2.4
mnormt	CRAN	1.5-5	1.5-5		1.5-4
modelr	CRAN	0.1.0			
modeltools	CRAN	0.2-21	0.2-21	0.2-21	
multtest	Bioconductor		2.30.0	2.30.0	
munsell	CRAN	0.4.3	0.4.3	0.4.3	0.4.3
mvtnorm	CRAN	1.0-5	1.0-5	1.0-6	1.0-5
network	CRAN		1.13.0		
nlme	base	3.1-128	3.1-128	3.1-131	3.1-128
nloptr	CRAN		1.0.4		1.0.4
NMF	CRAN	0.20.6	0.20.6	0.20.6	0.20.6
nnet	base	7.3-12	7.3-12	7.3-12	7.3-12
numDeriv	CRAN		2016.8-1		2014.2-1
openssl	CRAN	0.9.4	0.9.6	0.9.6	0.9.4

org.Hs.eg.db	Bioconductor		3.1.2		3.3.0
org.Sc.sgd.db	Bioconductor		3.4.0		
parallel	base	3.3.2	3.3.2	3.3.1	3.3.0
pathway.structure	GitHub	0.1.0	0.1.0	0.1.0	0.1.0
.permutation	TomKellyGenetics				
pbivnorm	CRAN		0.6.0		
PGSEA	Bioconductor		1.48.0		
pkgmaker	CRAN	0.22	0.22	0.22	0.22
PKI	CRAN		0.1-3		
plogr	CRAN		0.1-1	0.1-1	
plot.igraph	GitHub	0.0.0.9001	0.0.0.9001	0.0.0.9001	0.0.0.9001
	TomKellyGenetics				
plotrix	CRAN		3.6-4		
plyr	CRAN	1.8.4	1.8.4	1.8.4	1.8.3
png	CRAN		0.1-7		0.1-7
prabclus	CRAN	2.2-6	2.2-6	2.2-6	
praise	CRAN	1.0.0	1.0.0		1.0.0
pROC	CRAN		1.8	1.9.1	
prodlim	CRAN		1.5.7		
prof.tree	CRAN		0.1.0		
protools	CRAN		0.99-2		
progress	CRAN			1.1.2	
psych	CRAN	1.6.12	1.6.12		
purrr	CRAN	0.2.2	0.2.2	0.2.2	0.2.2
qgraph	CRAN		1.4.1		
quadprog	CRAN		1.5-5	1.5-5	1.5-5
R.methodsS3	CRAN		1.7.1		1.7.1
R.oo	CRAN		1.21.0		1.20.0
R.utils	CRAN		2.5.0		
R6	CRAN	2.1.3	2.2.0	2.2.0	2.1.3
RBGL	CRAN		1.50.0		
RColorBrewer	CRAN	1.1-2	1.1-2	1.1-2	1.1-2
Rcpp	CRAN	0.12.7	0.12.9	0.12.9	0.12.7
RcppArmadillo	CRAN			0.7.700.0.0	0.6.700.6.0
RcppEigen	CRAN		0.3.2.9.0		0.3.2.8.1

RCurl	CRAN		1.95-4.8	1.95-4.8	1.95-4.8
reactome.db	Bioconductor		1.52.1	1.52.1	
reactometree	GitHub		0.1		
	TomKellyGenetics				
readr	CRAN	1.0.0	1.0.0		
readxl	CRAN	0.1.1			
registry	CRAN	0.3	0.3	0.3	0.3
reshape2	CRAN	1.4.1	1.4.2	1.4.2	1.4.1
rgeff	CRAN		0.15.3	0.15.3	
rgl	CRAN			0.97.0	0.95.1441
Rgraphviz	CRAN		2.18.0		
rjson	CRAN		0.2.15		
RJSONIO	CRAN		1.3-0		
rmarkdown	CRAN		1.3	1.3	1
Rmpi	CRAN		0.6-6		0.6-5
rngtools	CRAN	1.2.4	1.2.4	1.2.4	1.2.4
robustbase	CRAN	0.92-7	0.92-7	0.92-7	0.92-5
ROCR	CRAN	1.0-7	1.0-7	1.0-7	1.0-7
Rook	CRAN		1.1-1	1.1-1	
roxygen2	CRAN	6.0.1	5.0.1	6.0.1	5.0.1
rpart	base	4.1-10	4.1-10	4.1-10	4.1-10
rprojroot	CRAN	1.2	1.1	1.2	
Rsamtools	Bioconductor		1.26.1	1.26.1	
rsconnect	CRAN		0.7		
RSQLite	CRAN		1.1-2	1.1-2	1.0.0
rstudioapi	CRAN	0.6	0.6	0.6	0.6
rvest	CRAN	0.3.2			
S4Vectors	Bioconductor		0.12.1	0.12.0	0.10.3
safe	Bioconductor		3.14.0	3.10.0	
scales	CRAN	0.4.0	0.4.1	0.4.1	0.4.0
selectr	CRAN	0.3-1			
sem	CRAN		3.1-8		
shiny	CRAN	0.14		1.0.0	
slipt	GitHub	0.1.0	0.1.0	0.1.0	0.1.0
	TomKellyGenetics				

sm	CRAN	2.2-5.4	2.2-5.4		
sna	CRAN		2.4		
snow	CRAN	0.4-1	0.4-2	0.4-2	0.3-13
sourcetools	CRAN	0.1.5		0.1.5	
SparseM	CRAN		1.74		1.7
spatial	base	7.3-11	7.3-11	7.3-11	7.3-11
splines	base	3.3.2	3.3.2	3.3.1	3.3.0
statnet.common	CRAN		3.3.0		
stats	base	3.3.2	3.3.2	3.3.1	3.3.0
stats4	base	3.3.2	3.3.2	3.3.1	3.3.0
stringi	CRAN	1.1.1	1.1.2	1.1.2	1.0-1
stringr	CRAN	1.1.0	1.1.0	1.2.0	1.0.0
Summarized Experiment	Bioconductor		1.4.0	1.4.0	
survival	base	2.39-4	2.40-1	2.40-1	2.39-4
tcltk	base	3.3.2	3.3.2	3.3.1	3.3.0
testthat	CRAN	1.0.2	1.0.2		1.0.2
tibble	CRAN	1.2	1.2	1.2	1.2
tidyverse	GitHub hadley	1.1.1			
timeline	CRAN		0.9		
tools	base	3.3.2	3.3.2	3.3.1	3.3.0
tpr	CRAN		0.3-1		
trimcluster	CRAN	0.1-2	0.1-2	0.1-2	
Unicode	CRAN	9.0.0-1	9.0.0-1	9.0.0-1	
utils	base	3.3.2	3.3.2	3.3.1	3.3.0
vioplot	CRAN		0.2		
vioplotx	GitHub TomKellyGenetics	0.0.0.9000	0.0.0.9000		
viridis	CRAN	0.3.4	0.3.4	0.3.4	
visNetwork	CRAN		1.0.3	1.0.3	
whisker	CRAN	0.3-2	0.3-2	0.3-2	0.3-2
withr	CRAN	1.0.2	1.0.2	1.0.2	1.0.2
XML	base	3.98-1.3	3.98-1.1	3.98-1.5	3.98-1.4

xml2	CRAN	1.1.1	1.1.1	1.0.0
xtable	CRAN	1.8-2	1.8-2	1.8-2
XVector	Bioconductor		0.14.0	0.14.0
yaml	CRAN		2.1.14	2.1.14
zlibbioc	CRAN		1.20.0	1.20.0
zoo	CRAN	1.7-13	1.7-14	1.7-13

Appendix C

Mutation Analysis in Breast Cancer

C.1 Synthetic Lethal Genes and Pathways

SLIPT expression analysis (described in Section 3.1) on TCGA breast cancer data ($n = 969$) found the following genes and pathways, described in sections 4.1 and 4.1.1.

Table C.1: Candidate synthetic lethal gene partners of *CDH1* from mtSLIPT

Gene	Observed	Expected	χ^2 value	p-value	p-value (FDR)
<i>TFAP2B</i>	8	36.7	89.5	3.60×10^{-20}	8.37×10^{-17}
<i>ZNF423</i>	15	36.7	78.8	7.89×10^{-18}	1.22×10^{-14}
<i>CALCOCO1</i>	11	36.7	76.8	2.09×10^{-17}	2.59×10^{-14}
<i>RBM5</i>	13	36.7	75.7	3.65×10^{-17}	4.00×10^{-14}
<i>BTG2</i>	7	36.7	71.7	2.72×10^{-16}	1.81×10^{-13}
<i>RXRA</i>	6	36.7	70.5	5.00×10^{-16}	2.97×10^{-13}
<i>SLC27A1</i>	11	36.7	70.3	5.42×10^{-16}	2.97×10^{-13}
<i>MEF2D</i>	12	36.7	69.6	7.86×10^{-16}	3.95×10^{-13}
<i>NISCH</i>	12	36.7	69.6	7.86×10^{-16}	3.95×10^{-13}
<i>AVPR2</i>	9	36.7	69.2	9.36×10^{-16}	4.58×10^{-13}
<i>CRY2</i>	13	36.7	68.9	1.07×10^{-15}	4.98×10^{-13}
<i>RAPGEF3</i>	13	36.7	68.9	1.07×10^{-15}	4.98×10^{-13}
<i>NRIP2</i>	10	36.7	68.2	1.58×10^{-15}	7.18×10^{-13}
<i>DARC</i>	12	36.7	66.4	3.76×10^{-15}	1.54×10^{-12}
<i>SFRS5</i>	12	36.7	66.4	3.76×10^{-15}	1.54×10^{-12}
<i>NOSTRIN</i>	5	36.7	65.1	7.40×10^{-15}	2.70×10^{-12}
<i>KIF13B</i>	12	36.7	63.4	1.69×10^{-14}	5.16×10^{-12}
<i>TENC1</i>	10	36.7	62.5	2.67×10^{-14}	7.40×10^{-12}
<i>MFAP4</i>	12	36.7	60.5	7.17×10^{-14}	1.67×10^{-11}
<i>ELN</i>	13	36.7	59.7	1.07×10^{-13}	2.32×10^{-11}
<i>SGK223</i>	14	36.7	59	1.51×10^{-13}	3.05×10^{-11}
<i>KIF12</i>	11	36.7	58.8	1.74×10^{-13}	3.34×10^{-11}
<i>SELP</i>	11	36.7	58.8	1.74×10^{-13}	3.34×10^{-11}
<i>CIRBP</i>	9	36.7	58.7	1.83×10^{-13}	3.41×10^{-11}
<i>CTDSP1</i>	9	36.7	58.7	1.83×10^{-13}	3.41×10^{-11}

Strongest candidate SL partners for *CDH1* by mtSLIPT with observed and expected numbers of *CDH1* mutant TCGA breast tumours with low expression of partner genes.

Table C.2: Pathways for *CDH1* partners from mtSLIPT

Pathways Over-represented	Pathway Size	SL Genes	p-value (FDR)
Eukaryotic Translation Elongation	86	60	2.0×10^{-128}
Peptide chain elongation	83	59	2.0×10^{-128}
Eukaryotic Translation Termination	83	58	2.3×10^{-125}
Viral mRNA Translation	81	57	2.5×10^{-124}
Nonsense Mediated Decay independent of the Exon Junction Complex	88	59	8.6×10^{-124}
Nonsense-Mediated Decay	103	61	5.2×10^{-117}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	103	61	5.2×10^{-117}
Formation of a pool of free 40S subunits	93	58	1.6×10^{-116}
L13a-mediated translational silencing of Ceruloplasmin expression	103	59	1.3×10^{-111}
3' -UTR-mediated translational regulation	103	59	1.3×10^{-111}
GTP hydrolysis and joining of the 60S ribosomal subunit	104	59	6.2×10^{-111}
SRP-dependent cotranslational protein targeting to membrane	104	58	2.9×10^{-108}
Eukaryotic Translation Initiation	111	59	3.0×10^{-106}
Cap-dependent Translation Initiation	111	59	3.0×10^{-106}
Influenza Viral RNA Transcription and Replication	108	57	5.1×10^{-103}
Influenza Infection	117	59	1.5×10^{-102}
Translation	141	64	3.7×10^{-101}
Influenza Life Cycle	112	57	1.4×10^{-100}
GPCR downstream signalling	472	116	1.0×10^{-80}
Hemostasis	422	105	1.4×10^{-78}

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in mtSLIPT partners for *CDH1*.

The genes and pathways identified in Tables C.1 and C.2 were derived from comparing the expression profiles of potential partners to the mutation status of *CDH1* (as shown in Figure 3.2). Thus the following analysis is only limited the samples for which TCGA provides both expression and somatic mutation data.

C.2 Synthetic Lethal Expression Profiles

Similar to the analysis of synthetic lethal partners against low *CDH1* expression in 4.1.2, the partners detected from *CDH1* mutation were also examined for their expression profiles and the pathway composition of gene clusters. Hierarchical clustering was performed on mtSLIPT partners for *CDH1* as showing in Figure C.1. Overrepresentation for Reactome pathways for each of the gene clusters identified is given in Table C.3.

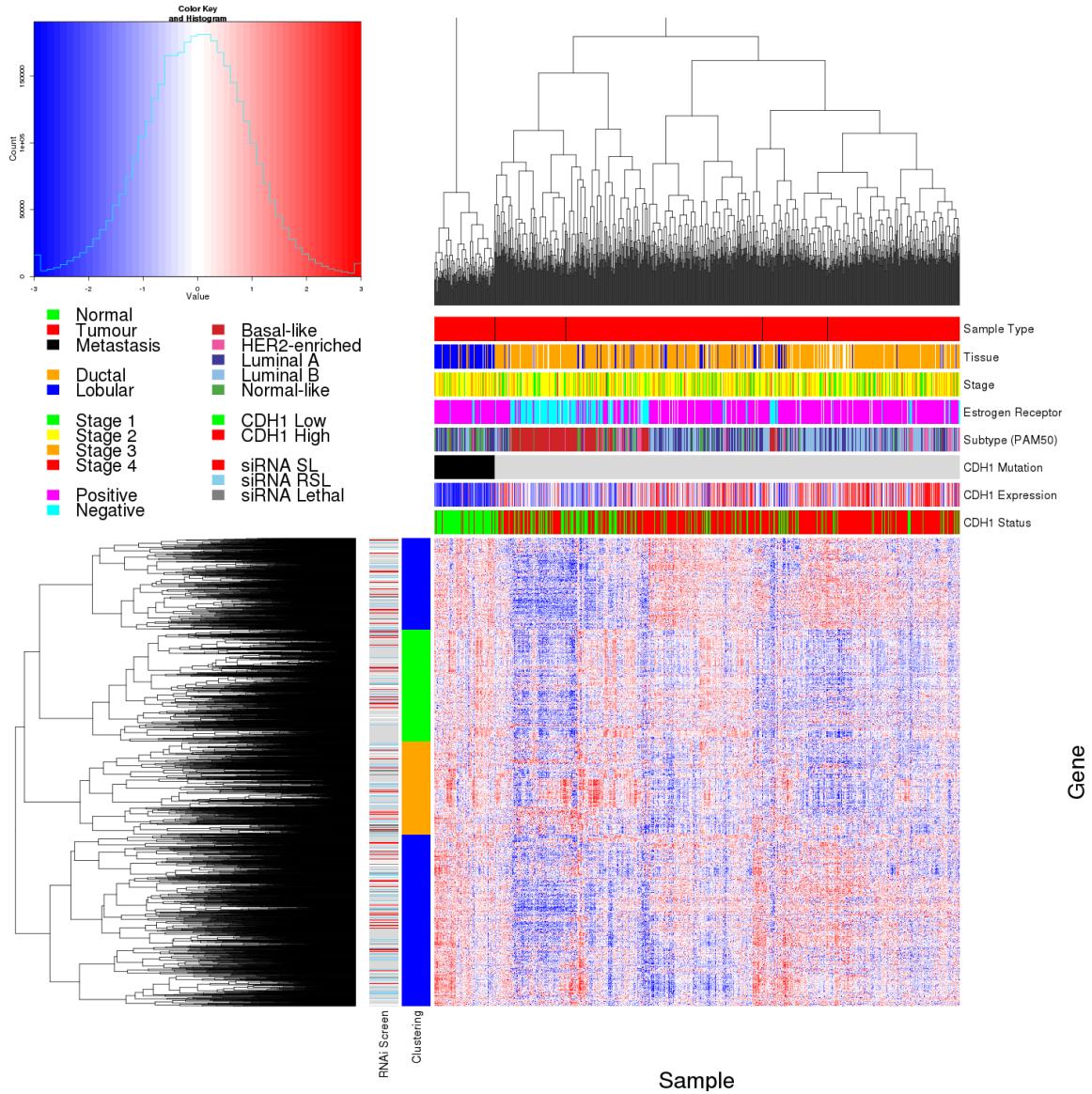


Figure C.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by *CDH1* somatic mutation status) analysed in TCGA breast cancer dataset for gene expression of 3,743 candidate partners of E-cadherin (*CDH1*) from mtSLIPT prediction (with significant FDR adjusted $p < 0.05$). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen Telford *et al.* (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in *CDH1* mutant samples and often lowly expressing *CDH1* wildtype samples (which were not tested for), although many of the *CDH1* mutant samples had among the lowest *CDH1* expression. In contrast to the expression analysis the (predominantly *CDH1* wildtype) basal subtype and estrogen receptor negative samples have depleted expression among most candidate synthetic lethal partners.

Table C.3: Pathway composition for clusters of *CDH1* partners from mtSLIPT

Pathways Over-represented in Cluster 1	Pathway Size	Cluster Genes	p-value (FDR)
Olfactory Signalling Pathway	57	8	7.1×10^{-9}
Assembly of the primary cilium	149	14	8.0×10^{-9}
Sphingolipid metabolism	62	8	9.6×10^{-9}
Signalling by ERBB4	133	12	5.1×10^{-8}
PI3K Cascade	65	7	4.9×10^{-7}
Circadian Clock	33	5	4.9×10^{-7}
Nuclear signalling by ERBB4	34	5	4.9×10^{-7}
Intraflagellar transport	35	5	4.9×10^{-7}
PI3K events in ERBB4 signalling	87	8	4.9×10^{-7}
PIP3 activates AKT signalling	87	8	4.9×10^{-7}
PI3K events in ERBB2 signalling	87	8	4.9×10^{-7}
PI-3K cascade:FGFR1	87	8	4.9×10^{-7}
PI-3K cascade:FGFR2	87	8	4.9×10^{-7}
PI-3K cascade:FGFR3	87	8	4.9×10^{-7}
PI-3K cascade:FGFR4	87	8	4.9×10^{-7}
Deadenylation of mRNA	22	4	5.6×10^{-7}
PI3K/AKT activation	90	8	5.6×10^{-7}
Cargo trafficking to the periciliary membrane	38	5	5.6×10^{-7}
Pathways Over-represented in Cluster 2	Pathway Size	Cluster Genes	p-value (FDR)
G _{αs} signalling events	83	19	5.1×10^{-25}
Extracellular matrix organization	238	30	1.4×10^{-18}
Hemostasis	422	46	2.7×10^{-16}
Aquaporin-mediated transport	32	9	2.7×10^{-16}
Transcriptional regulation of white adipocyte differentiation	56	11	1.7×10^{-15}
Degradation of the extracellular matrix	102	15	1.7×10^{-15}
Integration of energy metabolism	84	13	8.8×10^{-15}
GPCR downstream signalling	472	48	2.8×10^{-14}
G _{αz} signalling events	15	6	5.0×10^{-14}
Molecules associated with elastic fibres	33	8	5.4×10^{-14}
Phase 1 - Functionalization of compounds	67	11	5.6×10^{-14}
Platelet activation, signalling and aggregation	179	20	5.6×10^{-14}
Vasopressin regulates renal water homeostasis via Aquaporins	24	7	6.1×10^{-14}
Elastic fibre formation	37	8	$.03 \times 10^{-13}$
Calmodulin induced events	27	7	3.3×10^{-13}
CaM pathway	27	7	3.3×10^{-13}
cGMP effects	18	6	3.6×10^{-13}
G _{αi} signalling events	167	18	6.3×10^{-13}
Pathways Over-represented in Cluster 3	Pathway Size	Cluster Genes	p-value (FDR)
Eukaryotic Translation Elongation	86	55	1.1×10^{-112}
Peptide chain elongation	83	54	1.3×10^{-112}
Viral mRNA Translation	81	53	1.6×10^{-111}
Eukaryotic Translation Termination	83	53	7.1×10^{-110}
Nonsense Mediated Decay independent of the Exon Junction Complex	88	54	1.0×10^{-108}
Formation of a pool of free 40S subunits	93	53	4.1×10^{-102}
Nonsense-Mediated Decay	103	54	3.9×10^{-98}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	103	54	3.9×10^{-98}
L13a-mediated translational silencing of Ceruloplasmin expression	103	53	1.2×10^{-95}
3' -UTR-mediated translational regulation	103	53	1.2×10^{-95}
SRP-dependent cotranslational protein targeting to membrane	104	53	4.3×10^{-95}
GTP hydrolysis and joining of the 60S ribosomal subunit	104	53	4.3×10^{-95}
Influenza Viral RNA Transcription and Replication	108	53	9.6×10^{-93}
Eukaryotic Translation Initiation	111	53	4.2×10^{-91}
Cap-dependent Translation Initiation	111	53	4.2×10^{-91}
Influenza Life Cycle	112	53	1.4×10^{-90}
Influenza Infection	117	53	6.2×10^{-88}
Translation	141	55	3×10^{-81}
Pathways Over-represented in Cluster 4	Pathway Size	Cluster Genes	p-value (FDR)
ECM proteoglycans	66	10	2.9×10^{-11}
deactivation of the beta-catenin transactivating complex	38	7	5.1×10^{-10}
Arachidonic acid metabolism	41	7	1.1×10^{-9}
G _{αq} signalling events	149	14	4.0×10^{-9}
HS-GAG degradation	21	5	4.5×10^{-9}
Uptake and actions of bacterial toxins	22	5	6.1×10^{-9}
Gastrin-CREB signalling pathway via PKC and MAPK	170	15	6.1×10^{-9}
RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	64	8	6.1×10^{-9}
Non-integrin membrane-ECM interactions	53	7	1.5×10^{-8}
Syndecan interactions	25	5	1.5×10^{-8}
NOTCH1 Intracellular Domain Regulates Transcription	40	6	2.3×10^{-8}
Synthesis of Leukotrienes and Eoxins	15	4	3.2×10^{-8}
Signalling by NOTCH1	59	7	5.3×10^{-8}
Regulation of insulin secretion	44	6	6.0×10^{-8}
Metabolism of lipids and lipoproteins	471	37	8.2×10^{-8}
Signalling by NOTCH1	80	8	1.2×10^{-7}
Platelet activation, signalling and aggregation	179	14	1.2×10^{-7}
Recruitment of mitotic centrosome proteins and complexes	64	7	1.2×10^{-7}

Pathway over-representation analysis for Reactome pathways with the number of genes in each pathway (Pathway Size), number of genes within the pathway identified (Cluster Genes), and the pathway over-representation p-value (adjusted by FDR) from the hypergeometric test.

C.3 Comparison to Primary Screen

The mutation synthetic lethal partners with *CDH1* were also compared to siRNA primary screen data (Telford *et al.*, 2015), as performed in Section 4.2.1. These are expected to be more concordant with the experimental results performed on a null mutant, however this is not the case at the gene level: less genes overlapped with experimental candidates in Figure C.2. This may be affected by lower sample size for mutations in TCGA data or lower frequency (expected value) of *CDH1* mutations compared to low expression.

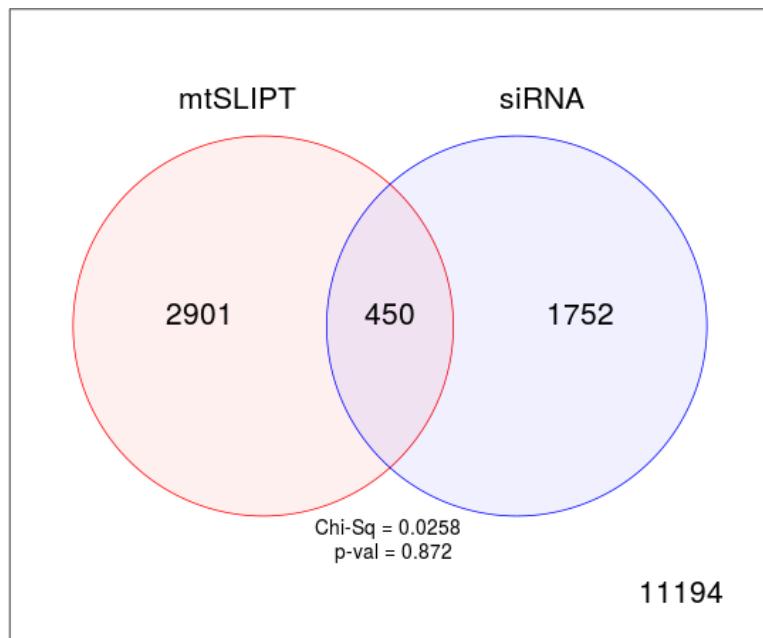


Figure C.2: **Comparison of mtSLIPT to siRNA.** Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ^2 test suggests that the overlap is no more than would be expected by chance ($p = 0.281$).

Despite a lower sample size (and low number of predicted partners) for mutation analysis, the pathway composition (Tables C.2 and C.4) is similar to expression analysis, as described in Section 4.2.5. In particular, the resampling analysis (Section C.3.1) supported many of the results of expression analysis (Section 4.2.5.1) with Tables C.5 and C.6 detecting many of the same or functionally-related pathways.

Table C.4: Pathway composition for *CDH1* partners from mtSLIPT and siRNA

Predicted only by SLIPT (2901 genes)	Pathway Size	Genes Identified	p-value (FDR)
Eukaryotic Translation Elongation	87	57	2.8×10^{-120}
Peptide chain elongation	84	56	3.1×10^{-120}
Eukaryotic Translation Termination	84	55	2.8×10^{-117}
Viral mRNA Translation	82	54	4.1×10^{-116}
Nonsense Mediated Decay independent of the Exon Junction Complex	89	55	3.7×10^{-113}
Formation of a pool of free 40S subunits	94	55	2.8×10^{-109}
Nonsense-Mediated Decay	104	57	8.4×10^{-108}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	104	57	8.4×10^{-108}
L13a-mediated translational silencing of Ceruloplasmin expression	104	56	3.4×10^{-105}
3' -UTR-mediated translational regulation	104	56	3.4×10^{-105}
GTP hydrolysis and joining of the 60S ribosomal subunit	105	56	1.4×10^{-104}
Eukaryotic Translation Initiation	112	56	2.8×10^{-100}
Cap-dependent Translation Initiation	112	56	2.8×10^{-100}
SRP-dependent cotranslational protein targeting to membrane	105	54	2.2×10^{-99}
Influenza Viral RNA Transcription and Replication	109	54	5.3×10^{-97}
Influenza Life Cycle	113	54	9.6×10^{-95}
Influenza Infection	118	55	1.7×10^{-94}
Translation	142	60	3.5×10^{-94}
Infectious disease	349	77	5.9×10^{-62}
Extracellular matrix organization	241	54	3.0×10^{-52}

Detected only by siRNA screen (1752 genes)	Pathway Size	Genes Identified	p-value (FDR)
Class A/1 (Rhodopsin-like receptors)	282	69	1.9×10^{-59}
GPCR ligand binding	363	78	2.7×10^{-54}
Peptide ligand-binding receptors	175	41	1.5×10^{-42}
$G_{\alpha i}$ signalling events	184	41	1.1×10^{-40}
Gastrin-CREB signalling pathway via PKC and MAPK	180	37	1.5×10^{-35}
$G_{\alpha q}$ signalling events	159	34	3.7×10^{-35}
DAP12 interactions	159	27	1.1×10^{-24}
VEGFA-VEGFR2 Pathway	91	19	1.0×10^{-23}
Downstream signal transduction	146	24	1.9×10^{-22}
Signalling by VEGF	99	19	2.6×10^{-22}
DAP12 signalling	149	24	4.2×10^{-22}
Organelle biogenesis and maintenance	264	34	4.3×10^{-20}
Downstream signalling of activated FGFR1	134	21	4.3×10^{-20}
Downstream signalling of activated FGFR2	134	21	4.3×10^{-20}
Downstream signalling of activated FGFR3	134	21	4.3×10^{-20}
Downstream signalling of activated FGFR4	134	21	4.3×10^{-20}
Signalling by ERBB2	146	22	5.3×10^{-20}
Signalling by FGFR	146	22	5.3×10^{-20}
Signalling by FGFR1	146	22	5.3×10^{-20}
Signalling by FGFR2	146	22	5.3×10^{-20}

Intersection of SLIPT and siRNA screen (450 genes)	Pathway Size	Genes Identified	p-value (FDR)
HS-GAG degradation	21	4	4.9×10^{-6}
Retinoid metabolism and transport	39	5	4.9×10^{-6}
Platelet activation, signalling and aggregation	186	13	4.9×10^{-6}
Signalling by NOTCH4	11	3	4.9×10^{-6}
$G_{\alpha s}$ signalling events	100	8	5.0×10^{-6}
Defective EXT2 causes exostoses 2	12	3	5.0×10^{-6}
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	12	3	5.0×10^{-6}
Class A/1 (Rhodopsin-like receptors)	289	18	2.2×10^{-5}
Signalling by PDGF	173	11	2.9×10^{-5}
Circadian Clock	34	4	2.9×10^{-5}
Signalling by ERBB4	139	9	4.3×10^{-5}
Role of LAT2/NTAL/LAB on calcium mobilization	99	7	4.4×10^{-5}
Peptide ligand-binding receptors	181	11	4.5×10^{-5}
Defective B4GALT7 causes EDS, progeroid type	19	3	4.5×10^{-5}
Defective B3GAT3 causes JDSSDHD	19	3	4.5×10^{-5}
Signalling by NOTCH	80	6	4.5×10^{-5}
$G_{\alpha q}$ signalling events	164	10	5.1×10^{-5}
Response to elevated platelet cytosolic Ca^{2+}	84	6	7.1×10^{-5}
Signalling by ERBB2	148	9	7.1×10^{-5}
Signalling by SCF-KIT	129	8	8.3×10^{-5}

C.3.1 Resampling Analysis

Table C.5: Pathways for *CDH1* partners from mtSLIPT

Reactome Pathway	Over-representation	Permutation
Eukaryotic Translation Elongation	3.2×10^{-128}	$< 7.035 \times 10^{-4}$
Peptide chain elongation	3.2×10^{-128}	$< 7.035 \times 10^{-4}$
Eukaryotic Translation Termination	3.7×10^{-125}	$< 7.035 \times 10^{-4}$
Viral mRNA Translation	4.1×10^{-124}	$< 7.035 \times 10^{-4}$
Nonsense Mediated Decay independent of the Exon Junction Complex	1.4×10^{-123}	$< 7.035 \times 10^{-4}$
Nonsense-Mediated Decay	8.4×10^{-117}	$< 7.035 \times 10^{-4}$
Nonsense Mediated Decay enhanced by the Exon Junction Complex	8.4×10^{-117}	$< 7.035 \times 10^{-4}$
Formation of a pool of free 40S subunits	2.6×10^{-116}	$< 7.035 \times 10^{-4}$
L13a-mediated translational silencing of Ceruloplasmin expression	2.0×10^{-111}	$< 7.035 \times 10^{-4}$
3' -UTR-mediated translational regulation	2.0×10^{-111}	$< 7.035 \times 10^{-4}$
GTP hydrolysis and joining of the 60S ribosomal subunit	9.9×10^{-111}	$< 7.035 \times 10^{-4}$
SRP-dependent cotranslational protein targeting to membrane	4.7×10^{-108}	$< 7.035 \times 10^{-4}$
Eukaryotic Translation Initiation	4.8×10^{-106}	$< 7.035 \times 10^{-4}$
Cap-dependent Translation Initiation	4.8×10^{-106}	$< 7.035 \times 10^{-4}$
Influenza Viral RNA Transcription and Replication	8.1×10^{-103}	$< 7.035 \times 10^{-4}$
Influenza Infection	2.4×10^{-102}	$< 7.035 \times 10^{-4}$
Translation	6.0×10^{-101}	$< 7.035 \times 10^{-4}$
Influenza Life Cycle	2.2×10^{-100}	$< 7.035 \times 10^{-4}$
Disease	2.1×10^{-90}	0.013347
GPCR downstream signalling	1.6×10^{-80}	0.095478
Hemostasis	2.1×10^{-78}	0.2671
Signalling by GPCR	1.2×10^{-73}	0.44939
<i>Extracellular matrix organization</i>	2.2×10^{-67}	0.054008
Metabolism of proteins	1.4×10^{-66}	0.9607
Signal Transduction	2.1×10^{-66}	0.48184
Developmental Biology	2.5×10^{-66}	0.54075
Innate Immune System	5.3×10^{-66}	0.9589
Infectious disease	9.6×10^{-66}	0.21075
Signalling by NGF	1.1×10^{-62}	0.43356
Immune System	2.8×10^{-62}	0.23052

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Table C.6: Pathways for *CDH1* partners from mtSLIPT and siRNA primary screen

Reactome Pathway	Over-representation	Permutation
Visual phototransduction	1.2×10^{-9}	0.86279
G_{as} signalling events	2.9×10^{-7}	0.023066
Retinoid metabolism and transport	2.9×10^{-7}	0.299
Acylic chain remodelling of PS	1.1×10^{-5}	0.42584
Transcriptional regulation of white adipocyte differentiation	1.1×10^{-5}	0.53928
Chemokine receptors bind chemokines	1.1×10^{-5}	0.95259
<i>Signalling by NOTCH4</i>	1.2×10^{-5}	0.079229
Defective EXT2 causes exostoses 2	1.2×10^{-5}	0.22292
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	1.2×10^{-5}	0.22292
Platelet activation, signalling and aggregation	1.2×10^{-5}	0.48853
Serotonin receptors	1.4×10^{-5}	0.34596
Nicotinamide salvaging	1.4×10^{-5}	0.70881
Phase 1 - Functionalization of compounds	2×10^{-5}	0.31142
Amine ligand-binding receptors	2.5×10^{-5}	0.34934
Acylic chain remodelling of PE	3.8×10^{-5}	0.42615
Signalling by GPCR	3.8×10^{-5}	0.93888
Molecules associated with elastic fibres	3.9×10^{-5}	0.017982
DAP12 interactions	3.9×10^{-5}	0.71983
Beta defensins	3.9×10^{-5}	0.91458
Cytochrome P ₄₅₀ - arranged by substrate type	4.7×10^{-5}	0.83493
GPCR ligand binding	5.7×10^{-5}	0.95258
Acylic chain remodelling of PC	6.1×10^{-5}	0.42584
Response to elevated platelet cytosolic Ca ²⁺	6.4×10^{-5}	0.54046
Arachidonic acid metabolism	6.7×10^{-5}	0.026696
Defective B4GALT7 causes EDS, progeroid type	7.3×10^{-5}	0.24921
Defective B3GAT3 causes JDSSDHD	7.3×10^{-5}	0.24921
Hydrolysis of LPC	7.3×10^{-5}	0.80663
Elastic fibre formation	7.4×10^{-5}	0.0058768
HS-GAG degradation	9.4×10^{-5}	0.0083179
<i>Bile acid and bile salt metabolism</i>	9.4×10^{-5}	0.079905
Netrin-1 signalling	0.00011	0.92216
Integration of energy metabolism	0.00011	0.011152
Dectin-2 family	0.00012	0.10385
Platelet sensitization by LDL	0.00012	0.34596
DAP12 signalling	0.00012	0.62787
Defensins	0.00012	0.77542
GPCR downstream signalling	0.00012	0.79454
<i>Diseases associated with glycosaminoglycan metabolism</i>	0.00013	0.065927
<i>Diseases of glycosylation</i>	0.00013	0.065927
Signalling by Retinoic Acid	0.00013	0.22292
Signalling by Leptin	0.00013	0.34596
Signalling by SCF-KIT	0.00013	0.70881
Opioid Signalling	0.00013	0.96053
Signalling by NOTCH	0.00015	0.26884
Platelet homeostasis	0.00015	0.4878
Signalling by NOTCH1	0.00016	0.13043
Class B/2 (Secretin family receptors)	0.00016	0.13994
<i>Diseases of Immune System</i>	0.0002	0.0795
<i>Diseases associated with the TLR signalling cascade</i>	0.0002	0.0795
A tetrasaccharide linker sequence is required for GAG synthesis	0.0002	0.42615

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

C.4 Compare SLIPT genes

The mutation synthetic lethal partners with *CDH1* were also compared to siRNA primary screen data (Telford *et al.*, 2015), by correlation and siRNA viability as described in sections 4.2.2 and 4.2.3.

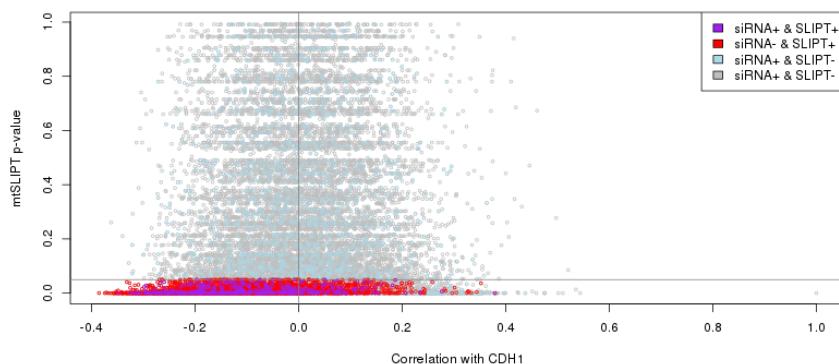


Figure C.3: Compare mtSLIPT and siRNA genes with correlation. The mtSLIPT p-values were compared against Pearson's correlation of expression with *CDH1*. Genes detected by SLIPT or siRNA are coloured according to the legend.

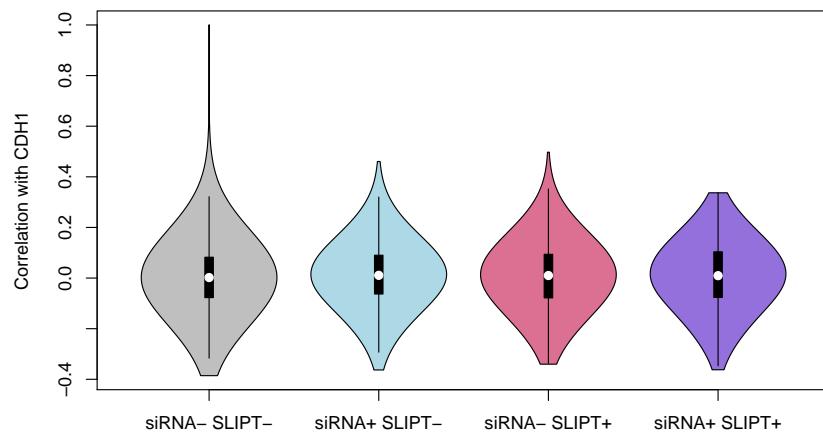


Figure C.4: Compare mtSLIPT and siRNA genes with correlation. Genes detected by mtSLIPT against *CDH1* mutation and siRNA screening were compared against Pearson's correlation of expression with *CDH1*. There were no differences in correlation between the gene groups.

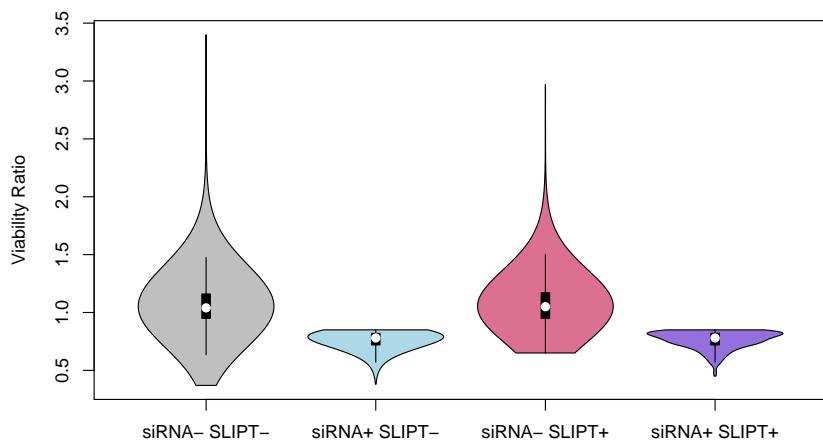


Figure C.5: Compare mtSLIPT and siRNA genes with siRNA viability. Genes detected as candidate synthetic lethal partners by mtSLIPT (in TCGA breast cancer) expression analysis against *CDH1* mutation and experimental screening (with siRNA) were compared against the viability ratio of *CDH1* mutant and wildtype cells in the primary siRNA screen. There were clear no differences in viability between genes detected by mtSLIPT and those not with the differences being primarily due to viability thresholds being used to detect synthetic lethality by Telford *et al.* (2015).

C.5 Metagene Analysis

Metagene analysis was also performed for synthetic lethal candidates for *CDH1* mutation. These are described and compared to expression analysis in Section 4.3.3.

Table C.7: Candidate synthetic lethal metagenes against *CDH1* from mtSLIPT

Pathway	ID	Observed	Expected	χ^2 value	p-value	p-value (FDR)
Neurotoxicity of clostridium toxins	168799	8	36.7	79.4	5.71×10^{-18}	3.14×10^{-15}
Aquaporin-mediated transport	445717	8	36.7	76.3	2.73×10^{-17}	9.01×10^{-15}
Toxicity of botulinum toxin type G (BoNT/G)	5250989	8	36.7	76.3	2.73×10^{-17}	9.01×10^{-15}
ABC-family proteins mediated transport	382556	10	36.7	68.2	1.58×10^{-15}	1.86×10^{-13}
G _{αz} signalling events	418597	10	36.7	59.9	9.97×10^{-14}	5.48×10^{-12}
Regulation of IGF transport and uptake by IGFBPs	381426	9	36.7	56.3	5.88×10^{-13}	2.11×10^{-11}
GP1b-IX-V activation signalling	430116	8	36.7	55.7	8.20×10^{-13}	2.76×10^{-11}
GABA receptor activation	977443	12	36.7	55.1	1.07×10^{-12}	3.26×10^{-11}
Vasopressin regulates renal water homeostasis via Aquaporins	432040	9	36.7	54.1	1.77×10^{-12}	4.88×10^{-11}
Toxicity of botulinum toxin type D (BoNT/D)	5250955	14	36.7	53.4	2.54×10^{-12}	6.64×10^{-11}
Toxicity of botulinum toxin type F (BoNT/F)	5250981	14	36.7	53.4	2.54×10^{-12}	6.64×10^{-11}
STAT6-mediated induction of chemokines	3249367	16	36.7	52.2	4.72×10^{-12}	1.13×10^{-10}
Toxicity of botulinum toxin type B (BoNT/B)	5250958	14	36.7	50.8	9.5×10^{-12}	1.98×10^{-10}
S6K1 signalling	165720	12	36.7	50.2	1.24×10^{-11}	2.5×10^{-10}
G _{αs} signalling events	418555	11	36.7	49.2	2.08×10^{-11}	3.85×10^{-10}
RHO GTPases activate CIT	5625900	14	36.7	48.2	3.34×10^{-11}	5.9×10^{-10}
NADE modulates death signalling	205025	15	36.7	47.4	5.00×10^{-11}	8.32×10^{-10}
Keratan sulfate degradation	2022857	10	36.7	46.6	7.5×10^{-11}	1.15×10^{-9}
Signalling by Retinoic Acid	5362517	10	36.7	46.6	7.5×10^{-11}	1.15×10^{-9}
Adenylate cyclase inhibitory pathway	170670	14	36.7	45.9	1.11×10^{-10}	1.59×10^{-9}
Inhibition of adenylate cyclase pathway	997269	14	36.7	45.9	1.11×10^{-10}	1.59×10^{-9}
Fatty acids	211935	6	36.7	45.7	1.21×10^{-10}	1.72×10^{-9}
Ionotropic activity of Kainate Receptors	451306	13	36.7	44.6	2.03×10^{-10}	2.58×10^{-9}
Activation of Ca-permeable Kainate Receptor	451308	13	36.7	44.6	2.03×10^{-10}	2.58×10^{-9}
RA biosynthesis pathway	5365859	13	36.7	44.6	2.03×10^{-10}	2.58×10^{-9}

Strongest candidate SL partners for *CDH1* by mtSLIPT with observed and expected numbers of mutant *CDH1* TCGA breast cancer tumours with low expression of partner metagenes.

C.6 Expression of Somatic Mutations

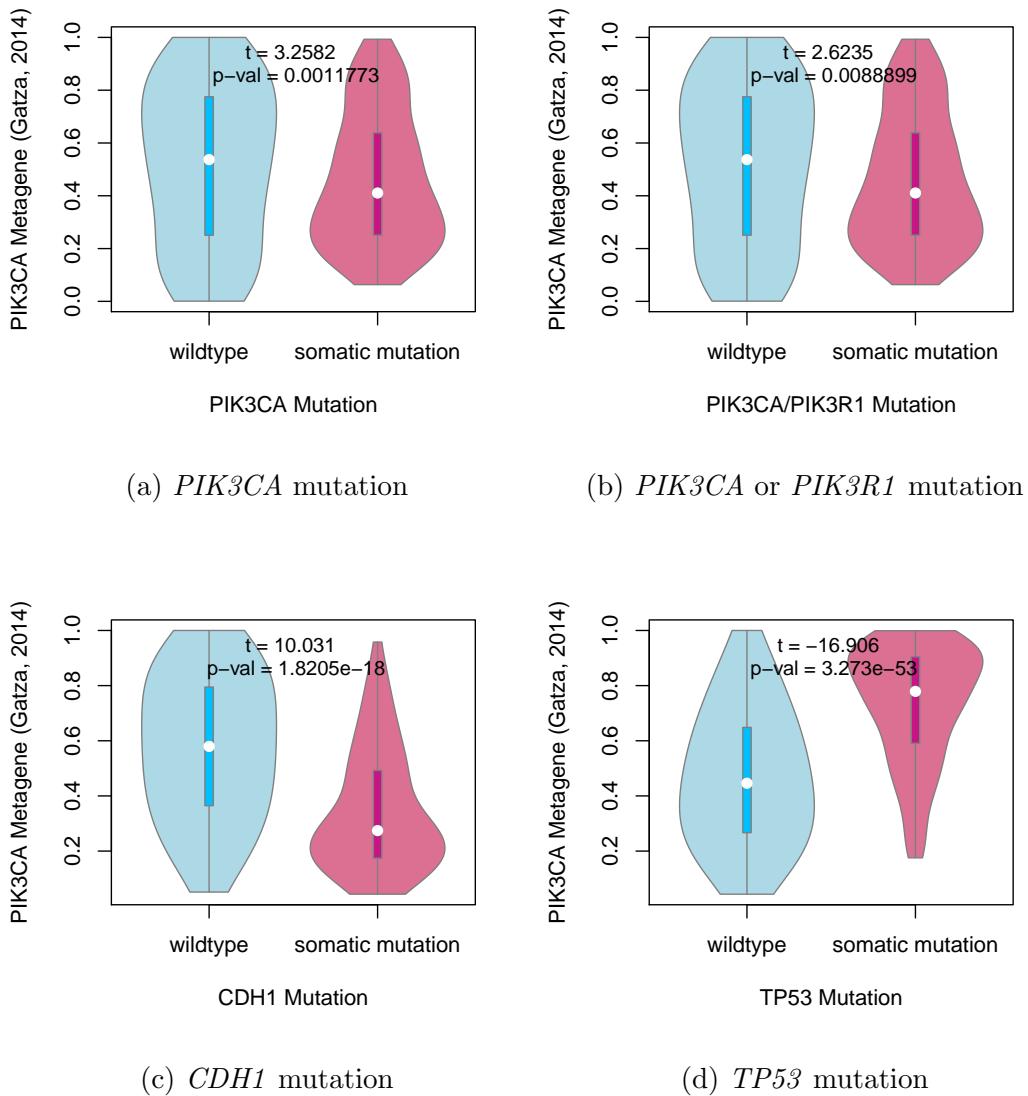


Figure C.6: **Somatic mutation against PIK3CA metagene.** Mutations in *PIK3CA*, *PIK3R1*, *CDH1*, and *TP53* were examined in TCGA breast cancer for their effect on the PIK3CA (Gatza *et al.*, 2014) pathway metagene. The tumour suppressors *CDH1* and *TP53* showed an increase and decrease in the metagene respectively, whereas *PIK3CA* and *PIK3R1* mutations weaker evidence of decrease in metagene levels.

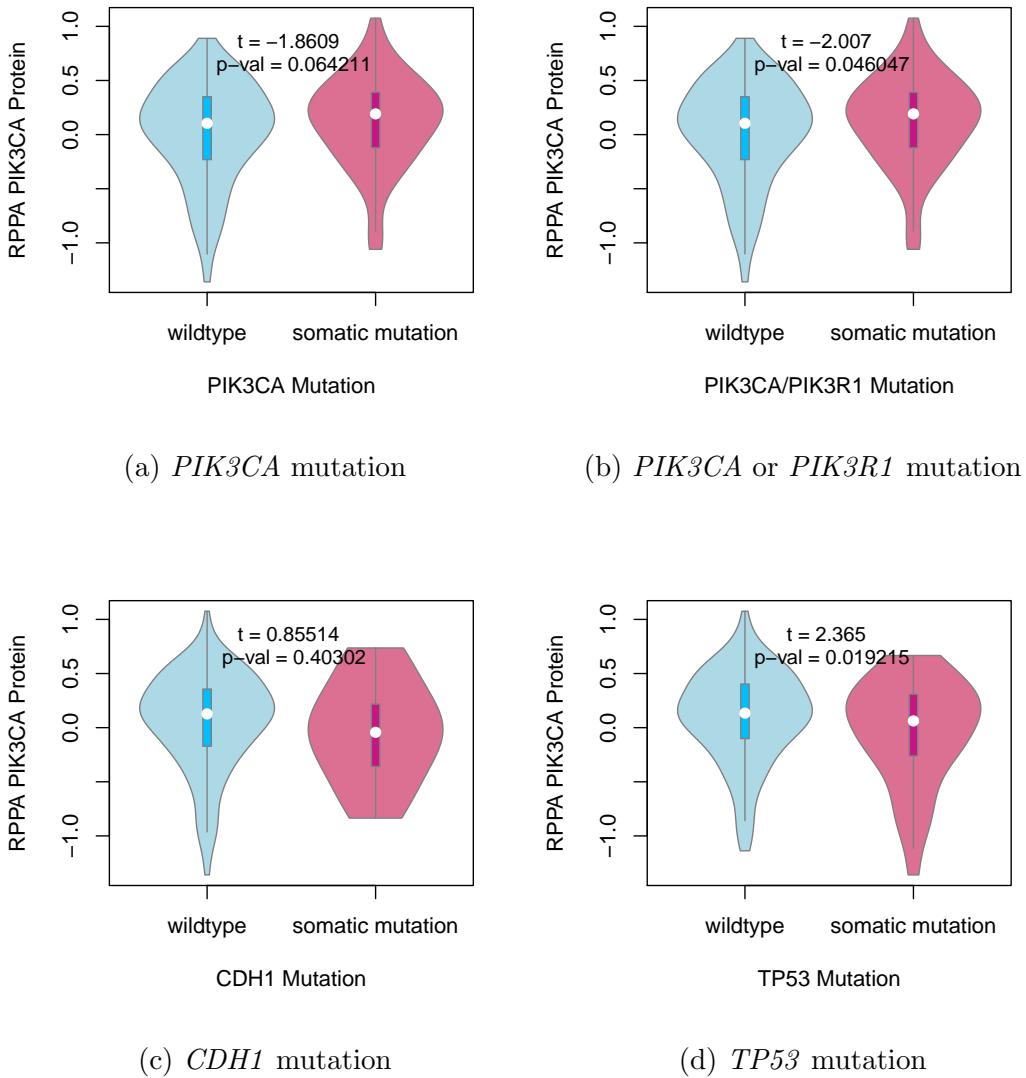


Figure C.7: Somatic mutation against PI3K protein. Mutations in *PIK3CA*, *PIK3R1*, *CDH1*, and *TP53* were examined in TCGA breast cancer for their effect on the expression of the p110 α protein (encoded by *PIK3CA*). Protein levels were significantly elevated in samples with *PIK3CA* or *PIK3R1* mutations and lower in samples with *TP53* mutations.

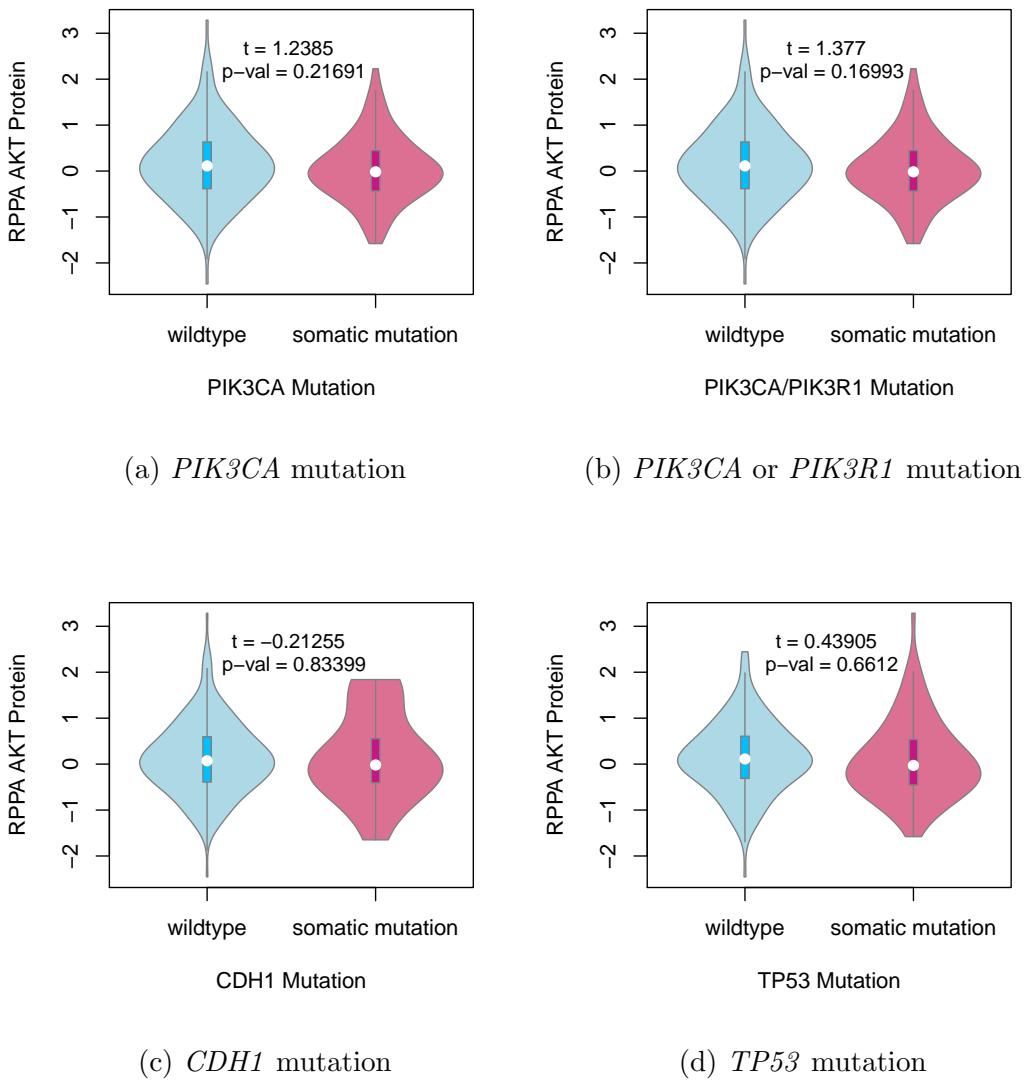


Figure C.8: Somatic mutation against AKT protein. Mutations in *PIK3CA*, *PIK3R1*, *CDH1*, and *TP53* were examined in TCGA breast cancer for their effect on the expression of the AKT protein (a downstream target of *PIK3CA*). Protein levels were not significantly different in samples mutations in any of these cancer genes.

C.7 Metagene Expression Profiles

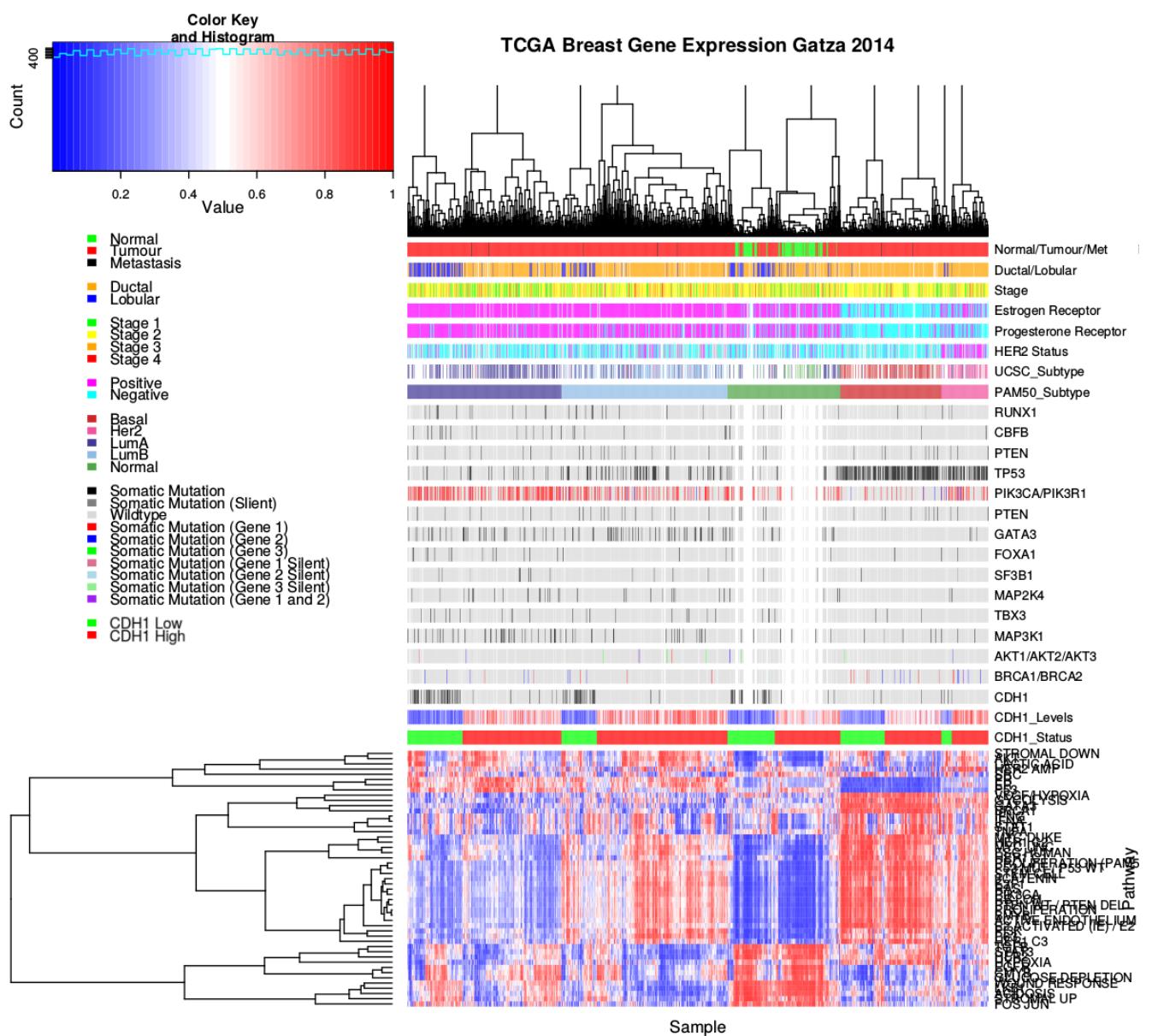


Figure C.9: **Pathway metagene expression profiles.** Expression profiles for metagene signatures from Gatza *et al.* (2014) in TCGA breast data, annotated for clinical factors and cancer gene mutations.

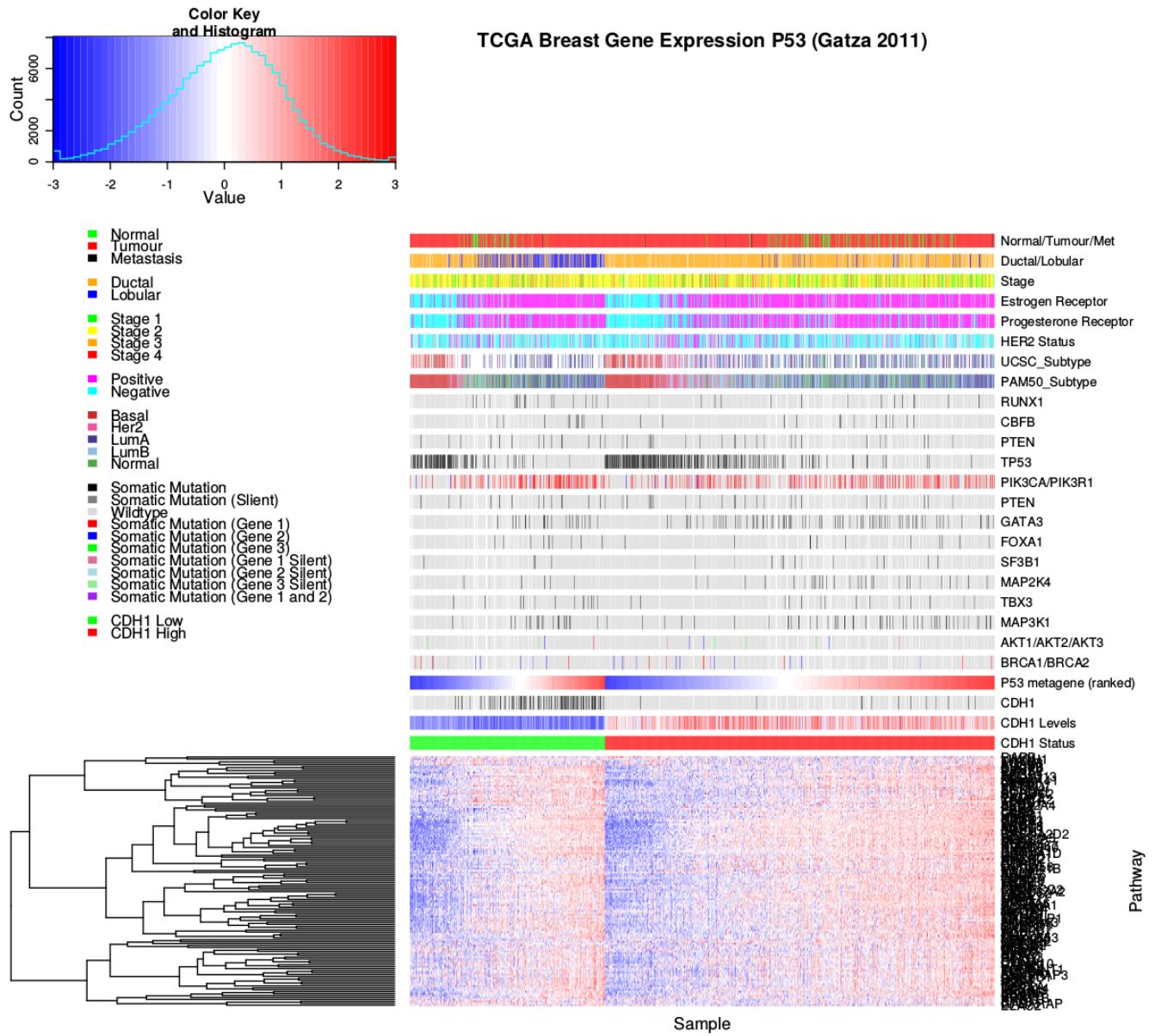


Figure C.10: Expression profiles for p53 related genes. Expression profiles the genes contained in the *TP53* gene signature from Gatza *et al.* (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by *CDH1* expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the metagene, with few very exceptions. *TP53* mutant samples had low metagene expression, consistent with loss of tumour suppressor functions, and were less likely to have *CDH1* or *PIK3CA* mutations.

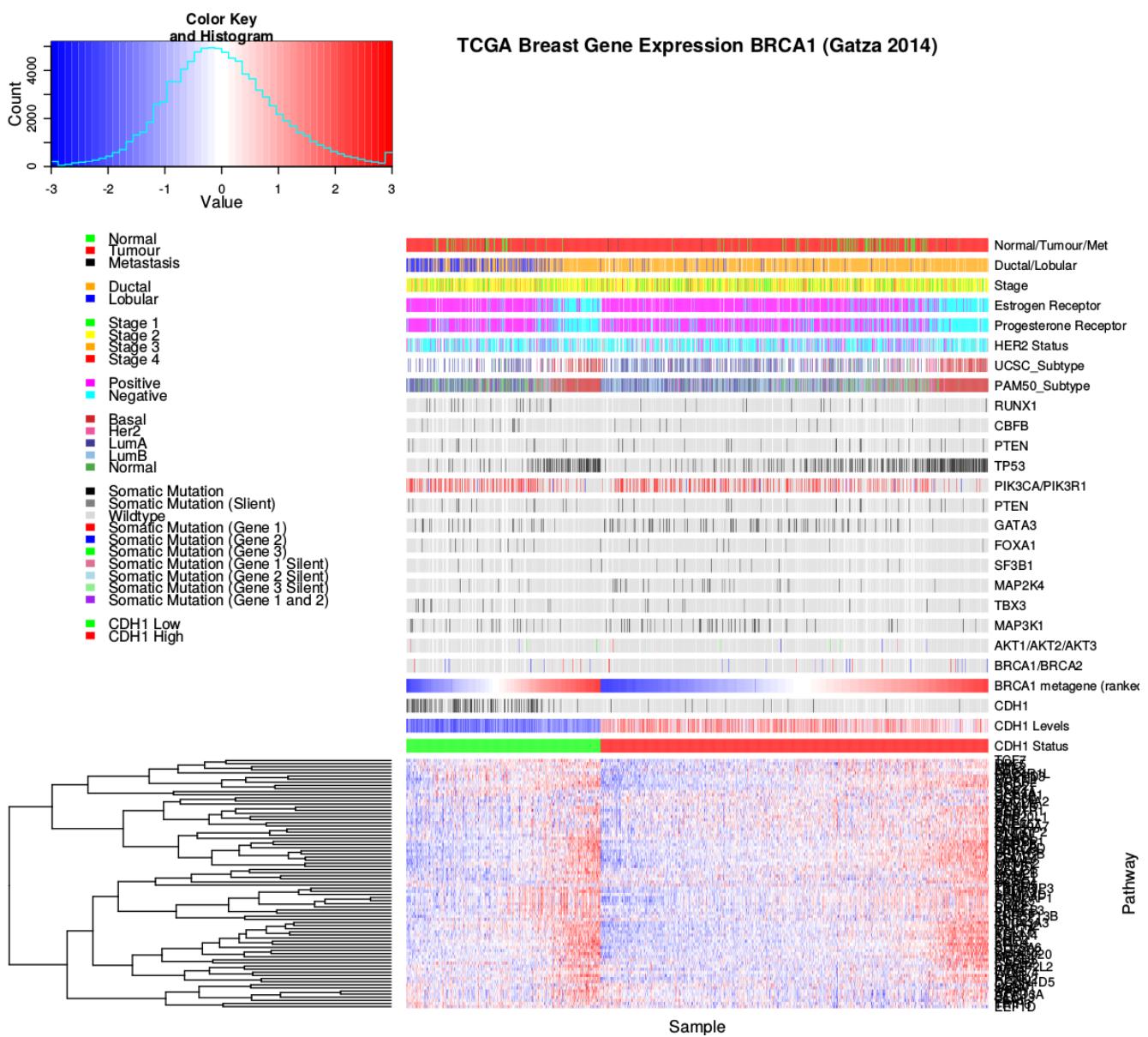


Figure C.11: **Expression profiles for BRCA related genes.** Expression profiles the genes contained in the gene signature related to *BRCA1* and *BRCA2* functions from Gatza *et al.* (2014) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by *CDH1* expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the metagene, with few very exceptions. *BRCA1* and *BRCA2* mutant samples had higher metagene expression than most samples for the ductal subtype, although this was not the case (for the lobular samples for which the metagene was lower). However, the metagene was higher for basal subtype and estrogen receptor negative samples.

Appendix D

Intrinsic Subtyping

The intrinsic subtypes for TCGA breast cancer samples provided by UCSC (TCGA, 2012) that were derived from microarray analysis have been compared to the PAM50 results for performing subtyping from RNA-Seq data (Parker *et al.*, 2009). As shown in Table D.1, these subtypes were highly concordant for samples which had both procedures performed upon them ($\chi^2 = 1305.9$, $p = 2.73 \times 10^{-268}$). The main exception were the luminal A samples some of which were reclassified as luminal B or “normal-like”.

Table D.1: Comparison of Intrinsic Subtypes

UCSC Subtype					
	Basal-like	HER2-enriched	Luminal A	Luminal B	Normal-like
	100	58	232	128	30
PAM50 Subtype					
	Basal-like	HER2-enriched	Luminal A	Luminal B	Normal-like
	208	94	314	334	227
UCSC Subtype					
PAM50 Subtype	Basal-like	HER2-enriched	Luminal A	Luminal B	Normal-like
Basal-like	96	4	2	2	1
HER2-enriched	0	47	5	3	0
Luminal A	1	0	141	1	0
Luminal B	2	7	49	121	0
Normal-like	1	0	35	1	29

The intrinsic subtypes of TCGA breast samples were compared between those provided by UCSC (TCGA, 2012) from microarray expression to those derived from RNA-Seq data (Parker *et al.*, 2009). Comparisons between these were limited to samples for which both data types were available.

The PAM50 subtypes are potentially more accurate given similarity of these subtypes and that the remainder of the subtypes were accurately recapitulated with RNA-

Seq data. Furthermore, UCSC subtypes correctly identified 22/22 normal samples as “normal-like” and PAM50 subtyping in RNA-Seq data had a success rate of 112/113 (including all of those identified from microarrays). Therefore the PAM50 subtypes (performed on a larger cohort of samples) are appropriate to use for further interpretation, superceeding the UCSC subtypes available for a limited set of samples.

Appendix E

Stomach Expression Analysis

The following results are a replication of the TCGA results (in Chapter 4) with stomach cancer data, using synthetic lethality (SLIPT) against *CDH1*.

E.1 Synthetic Lethal Genes and Pathways

Table E.1: Synthetic lethal gene partners of *CDH1* from SLIPT in stomach cancer

Gene	Observed	Expected	χ^2 value	p-value	p-value (FDR)
<i>PRAF2</i>	17	50.4	121	3.54×10^{-25}	1.45×10^{-21}
<i>EMP3</i>	17	50.4	115	5.06×10^{-24}	1.48×10^{-20}
<i>PLEKHO1</i>	22	50.4	112	2.14×10^{-23}	4.75×10^{-20}
<i>SELM</i>	20	50.4	111	5.13×10^{-23}	8.09×10^{-20}
<i>GYPC</i>	20	50.4	110	5.77×10^{-23}	8.45×10^{-20}
<i>COX7A1</i>	18	50.4	109	1.15×10^{-22}	1.39×10^{-19}
<i>TNFSF12</i>	20	50.4	106	4.06×10^{-22}	4.38×10^{-19}
<i>SEPT4</i>	17	50.4	106	6.58×10^{-22}	5.91×10^{-19}
<i>LGALS1</i>	19	50.4	105	6.64×10^{-22}	5.91×10^{-19}
<i>RARRES2</i>	27	50.4	105	8.02×10^{-22}	6.85×10^{-19}
<i>VEGFB</i>	16	50.4	104	1.19×10^{-21}	9.74×10^{-19}
<i>PRR24</i>	22	50.4	102	2.96×10^{-21}	2.02×10^{-18}
<i>SYNC</i>	19	50.4	102	3.73×10^{-21}	2.39×10^{-18}
<i>MAGEH1</i>	17	50.4	100	9.52×10^{-21}	5.01×10^{-18}
<i>HSPB2</i>	23	50.4	99.6	1.19×10^{-20}	5.82×10^{-18}
<i>SMARCD3</i>	19	50.4	99	1.59×10^{-20}	7.57×10^{-18}
<i>CREM</i>	13	50.4	98.1	2.48×10^{-20}	1.13×10^{-17}
<i>GNG11</i>	20	50.4	97.3	3.68×10^{-20}	1.59×10^{-17}
<i>GNAI2</i>	17	50.4	96.4	5.75×10^{-20}	2.36×10^{-17}
<i>FUNDC2</i>	22	50.4	95.9	7.39×10^{-20}	2.91×10^{-17}
<i>CNRIP1</i>	21	50.4	95.3	1.0×10^{-19}	3.66×10^{-17}
<i>CALHM2</i>	22	50.4	93.1	2.94×10^{-19}	1.06×10^{-16}
<i>ARID5A</i>	18	50.4	92.7	3.47×10^{-19}	1.22×10^{-16}
<i>ST3GAL3</i>	27	50.4	92.2	4.49×10^{-19}	1.56×10^{-16}
<i>LOC339524</i>	21	50.4	92.1	4.8×10^{-19}	1.59×10^{-16}

SLIPT partners of *CDH1* with observed and expected numbers of TCGA stomach cancer samples with low expression of both genes.

Table E.2: Pathways for *CDH1* partners from SLIPT in stomach cancer

Pathways Over-represented	Pathway Size	SL Genes	p-value (FDR)
Extracellular matrix organization	241	104	7.5×10^{-140}
Hemostasis	445	138	1.8×10^{-121}
Developmental Biology	432	125	9.2×10^{-107}
Axon guidance	289	94	1.5×10^{-102}
Eukaryotic Translation Termination	84	49	1.9×10^{-99}
GPCR ligand binding	373	108	3.8×10^{-99}
Viral mRNA Translation	82	48	3.3×10^{-98}
Formation of a pool of free 40S subunits	94	51	3.3×10^{-98}
Eukaryotic Translation Elongation	87	49	1.6×10^{-97}
Peptide chain elongation	84	48	7.2×10^{-97}
Class A/1 (Rhodopsin-like receptors)	289	90	2.7×10^{-96}
Nonsense Mediated Decay independent of the Exon Junction Complex	89	49	3.0×10^{-96}
Infectious disease	349	100	2.6×10^{-94}
GTP hydrolysis and joining of the 60S ribosomal subunit	105	52	3.4×10^{-94}
L13a-mediated translational silencing of Ceruloplasmin expression	104	51	2.8×10^{-92}
3' -UTR-mediated translational regulation	104	51	2.8×10^{-92}
Neuronal System	272	84	8.4×10^{-92}
SRP-dependent cotranslational protein targeting to membrane	105	51	9.5×10^{-92}
Eukaryotic Translation Initiation	112	52	2.0×10^{-90}
Cap-dependent Translation Initiation	112	52	2.0×10^{-90}

Gene set over-representation analysis (hypergeometric test) for Reactome pathways for *CDH1*.

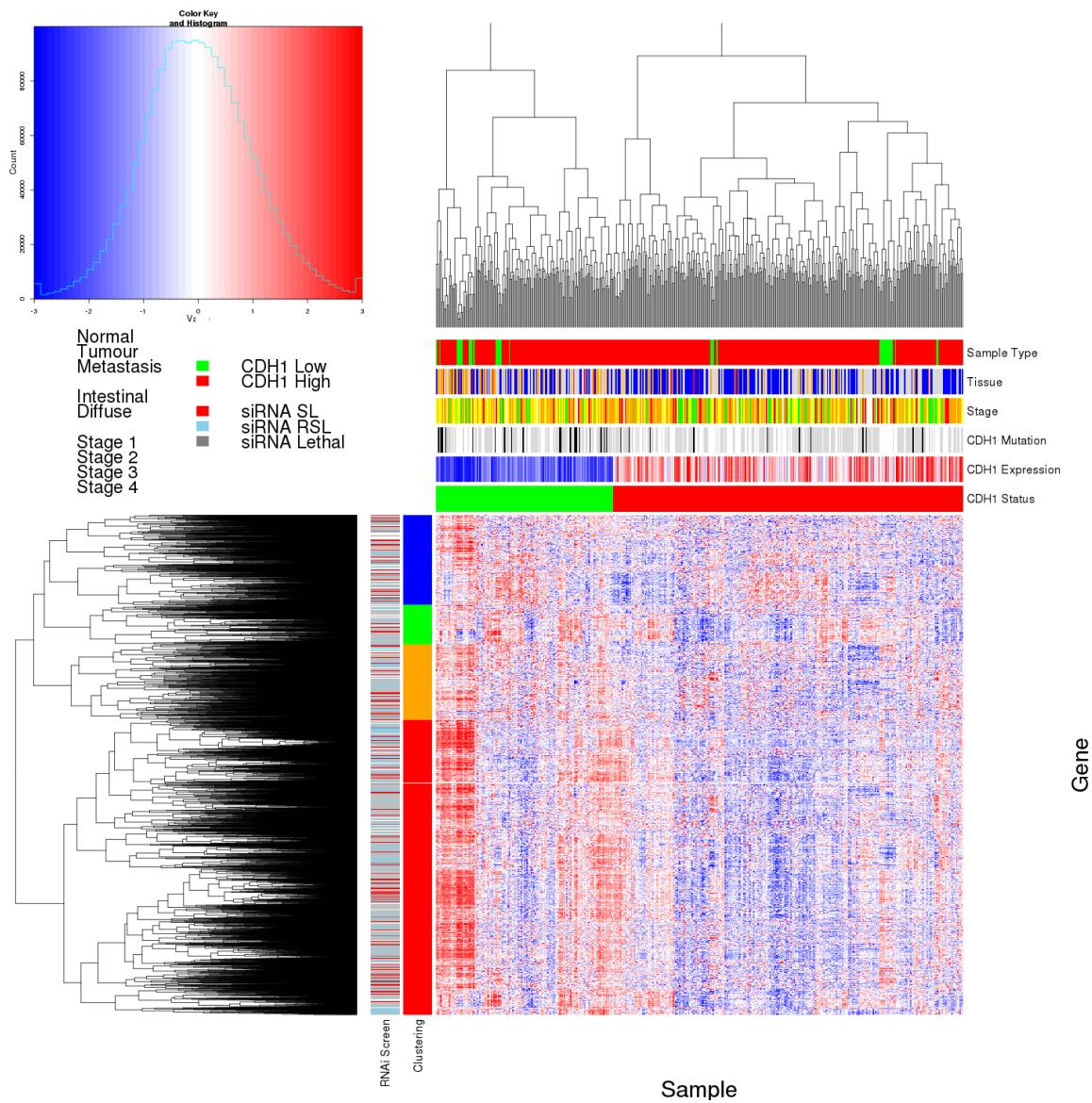


Figure E.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by the $1/3$ quantile of *CDH1* expression) analysed in TCGA stomach cancer dataset for gene expression of 4,365 candidate partners of E-cadherin (*CDH1*) from SLIPT prediction (with significant FDR adjusted $p < 0.05$). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen Telford *et al.* (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in *CDH1* low samples, notably diffuse and *CDH1* mutant samples have elevated expression in one or more distinct clusters, although there was less complexity and variation among candidate synthetic lethal partners than in breast data. *CDH1* low samples also contained most of samples with *CDH1* mutations.

Table E.3: Pathway composition for clusters of *CDH1* partners in stomach SLIPT

Pathways Over-represented in Cluster 1	Pathway Size	Cluster Genes	p-value (FDR)
Viral mRNA Translation	82	48	1.3×10^{-97}
Formation of a pool of free 40S subunits	94	51	1.3×10^{-97}
Eukaryotic Translation Elongation	87	49	4.8×10^{-97}
Peptide chain elongation	84	48	1.4×10^{-96}
Eukaryotic Translation Termination	84	48	1.4×10^{-96}
GTP hydrolysis and joining of the 60S ribosomal subunit	105	52	7.9×10^{-94}
Nonsense Mediated Decay independent of the Exon Junction Complex	89	48	3.1×10^{-93}
L13a-mediated translational silencing of Ceruloplasmin expression	104	51	5.1×10^{-92}
3' -UTR-mediated translational regulation	104	51	5.1×10^{-92}
SRP-dependent cotranslational protein targeting to membrane	105	51	1.7×10^{-91}
Eukaryotic Translation Initiation	112	52	3.3×10^{-90}
Cap-dependent Translation Initiation	112	52	3.3×10^{-90}
Translation	142	56	3.6×10^{-85}
Nonsense-Mediated Decay	104	48	1.2×10^{-84}
Nonsense Mediated Decay enhanced by the Exon Junction Complex	104	48	1.2×10^{-84}
Influenza Viral RNA Transcription and Replication	109	48	4.1×10^{-82}
Influenza Life Cycle	113	48	3.4×10^{-80}
Influenza Infection	118	48	6.4×10^{-78}
Pathways Over-represented in Cluster 2	Pathway Size	Cluster Genes	p-value (FDR)
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	65	12	1.3×10^{-15}
Phosphorylation of CD3 and TCR zeta chains	18	6	1.7×10^{-12}
Generation of second messenger molecules	29	7	2.7×10^{-12}
PD-1 signalling	21	6	7.4×10^{-12}
TCR signalling	62	9	4.3×10^{-11}
Translocation of ZAP-70 to Immunological synapse	16	5	1.1×10^{-10}
Interferon alpha/beta signalling	68	9	1.6×10^{-10}
Initial triggering of complement	17	5	1.6×10^{-10}
IKK complex recruitment mediated by RIP1	19	5	5.1×10^{-10}
TRIF-mediated programmed cell death	10	4	6.2×10^{-10}
Creation of C4 and C2 activators	11	4	1.3×10^{-9}
RHO GTPases Activate NADPH Oxidases	11	4	1.3×10^{-9}
Interferon Signalling	175	15	2.3×10^{-9}
Chemokine receptors bind chemokines	52	7	4.0×10^{-9}
Interferon gamma signalling	74	8	1.6×10^{-8}
TRAF6 mediated induction of TAK1 complex	15	4	1.6×10^{-8}
Activation of IRF3/IRF7 mediated by TBK1/IKK epsilon	16	4	2.7×10^{-8}
Downstream TCR signalling	45	6	3.5×10^{-8}
Pathways Over-represented in Cluster 3	Pathway Size	Cluster Genes	p-value (FDR)
Uptake and actions of bacterial toxins	22	4	3.5×10^{-6}
Neurotoxicity of clostridium toxins	10	3	3.5×10^{-6}
Activation of PPARGC1A (PGC-1alpha) by phosphorylation	10	3	3.5×10^{-6}
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	28	4	1.4×10^{-5}
Assembly of the primary cilium	149	10	2.5×10^{-5}
Serotonin Neurotransmitter Release Cycle	15	3	2.5×10^{-5}
Glycosaminoglycan metabolism	114	8	3.3×10^{-5}
Platelet homeostasis	54	5	3.3×10^{-5}
Norepinephrine Neurotransmitter Release Cycle	17	3	3.3×10^{-5}
Acetylcholine Neurotransmitter Release Cycle	17	3	3.3×10^{-5}
G _{αs} signalling events	100	7	5.5×10^{-5}
GABA synthesis, release, reuptake and degradation	19	3	5.6×10^{-5}
deactivation of the beta-catenin transactivating complex	39	4	6.7×10^{-5}
Dopamine Neurotransmitter Release Cycle	20	3	6.7×10^{-5}
IRS-related events triggered by IGF1R	83	6	7.1×10^{-5}
Generic Transcription Pathway	186	11	7.1×10^{-5}
Termination of O-glycan biosynthesis	21	3	7.4×10^{-5}
Kinesins	22	3	8.5×10^{-5}
Pathways Over-represented in Cluster 4	Pathway Size	Cluster Genes	p-value (FDR)
Extracellular matrix organization	241	97	8.8×10^{-126}
Axon guidance	289	75	8.3×10^{-72}
Hemostasis	445	101	8.3×10^{-72}
Developmental Biology	432	95	3.0×10^{-67}
Response to elevated platelet cytosolic Ca ²⁺	84	37	5.8×10^{-67}
Platelet degranulation	79	36	5.8×10^{-67}
Degradation of the extracellular matrix	104	39	6.7×10^{-63}
Platelet activation, signalling and aggregation	186	52	6.6×10^{-62}
ECM proteoglycans	66	31	8.1×10^{-61}
Neuronal System	272	64	5.1×10^{-60}
Signalling by PDGF	173	47	9.7×10^{-57}
Integrin cell surface interactions	82	31	1.9×10^{-53}
Collagen biosynthesis and modifying enzymes	56	26	1.1×10^{-52}
Collagen formation	67	28	1.4×10^{-52}
Class A/1 (Rhodopsin-like receptors)	289	61	2.3×10^{-52}
GPCR ligand binding	373	73	2.8×10^{-52}
Elastic fibre formation	38	22	4.7×10^{-52}
Non-integrin membrane-ECM interactions	53	24	7.0×10^{-49}

Pathway over-representation analysis for Reactome pathways with the number of genes in each pathway (Pathway Size), number of genes within the pathway identified (Cluster Genes), and the pathway over-representation p-value (adjusted by FDR) from the hypergeometric test.

E.2 Comparison to Primary Screen

The synthetic lethal partners with *CDH1* expression in stomach cancers were also compared to siRNA primary screen data (Telford *et al.*, 2015), as performed in Section 4.2.1. These are expected to be more concordant with the experimental results performed on a null mutant, however this is not the case at the gene level: less genes overlapped with experimental candidates in Figure E.2. This may be affected by lower sample size for mutations in TCGA data or lower frequency (expected value) of *CDH1* mutations compared to low expression.

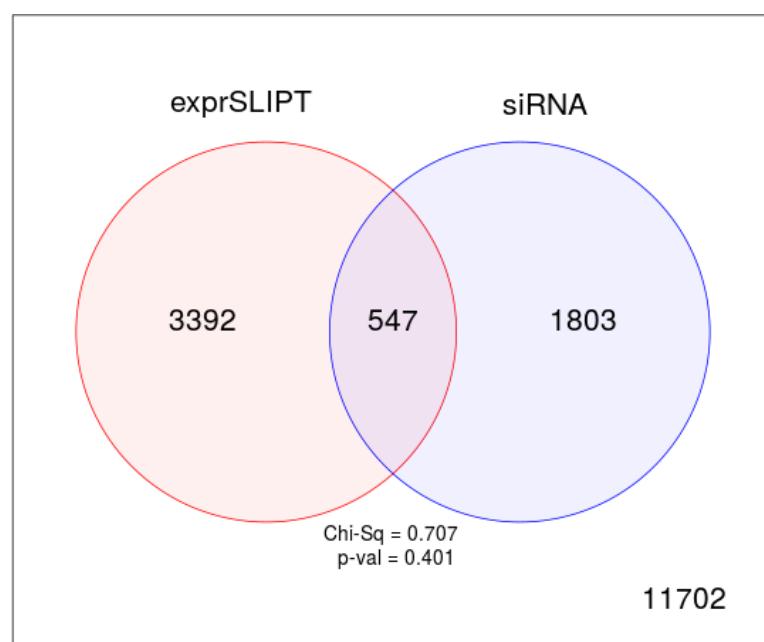


Figure E.2: **Comparison of SLIPT in stomach to siRNA.** Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ^2 test suggests that the overlap is no more than would be expected by chance ($p = 0.281$).

Table E.4: Pathway composition for *CDH1* partners from SLIPT and siRNA screening

Predicted only by SLIPT (3392 genes)	Pathway	Size	Genes Identified	p-value (FDR)
Extracellular matrix organization		238	90	3.4×10^{-107}
Eukaryotic Translation Termination		79	46	7.6×10^{-91}
Viral mRNA Translation		77	45	1.2×10^{-89}
Eukaryotic Translation Elongation		82	46	5.8×10^{-89}
Peptide chain elongation		79	45	2.1×10^{-88}
Nonsense Mediated Decay independent of the Exon Junction Complex		84	46	9.4×10^{-88}
Formation of a pool of free 40S subunits		89	47	3.3×10^{-87}
GTP hydrolysis and joining of the 60S ribosomal subunit		100	48	3.2×10^{-83}
Axon guidance		284	84	3.9×10^{-82}
Developmental Biology		426	111	4.2×10^{-82}
L13a-mediated translational silencing of Ceruloplasmin expression		99	47	1.4×10^{-81}
3' -UTR-mediated translational regulation		99	47	1.4×10^{-81}
SRP-dependent cotranslational protein targeting to membrane		99	47	1.4×10^{-81}
Nonsense-Mediated Decay		99	47	1.4×10^{-81}
Nonsense Mediated Decay enhanced by the Exon Junction Complex		99	47	1.4×10^{-81}
Hemostasis		438	112	1.2×10^{-80}
Eukaryotic Translation Initiation		107	48	8.0×10^{-80}
Cap-dependent Translation Initiation		107	48	8.0×10^{-80}
Infectious disease		338	90	1.6×10^{-76}
Neuronal System		267	77	1.6×10^{-76}
Detected only by siRNA screen (1803 genes)	Pathway	Size	Genes Identified	p-value (FDR)
Class A/1 (Rhodopsin-like receptors)		282	62	8.1×10^{-50}
GPCR ligand binding		363	71	4.9×10^{-46}
Peptide ligand-binding receptors		175	38	7.9×10^{-38}
G _{αi} signalling events		184	37	1.1×10^{-34}
Gastrin-CREB signalling pathway via PKC and MAPK		180	35	1.4×10^{-32}
G _{αq} signalling events		159	32	4.8×10^{-32}
DAP12 interactions		159	29	1.4×10^{-27}
Downstream signal transduction		146	26	2.4×10^{-25}
DAP12 signalling		149	26	6.4×10^{-25}
VEGFA-VEGFR2 Pathway		91	19	8.1×10^{-24}
Signalling by PDGF		172	27	5.7×10^{-23}
Signalling by ERBB2		146	24	1.4×10^{-22}
Signalling by VEGF		99	19	2.0×10^{-22}
Visual phototransduction		85	17	1.3×10^{-21}
Downstream signalling of activated FGFR1		134	22	1.3×10^{-21}
Downstream signalling of activated FGFR2		134	22	1.3×10^{-21}
Downstream signalling of activated FGFR3		134	22	1.3×10^{-21}
Downstream signalling of activated FGFR4		134	22	1.3×10^{-21}
Signalling by FGFR		146	23	2.0×10^{-21}
Signalling by FGFR1		146	23	2.0×10^{-21}
Intersection of SLIPT and siRNA screen (547 genes)	Pathway	Size	Genes Identified	p-value (FDR)
Class A/1 (Rhodopsin-like receptors)		282	25	3.9×10^{-9}
Platelet activation, signalling and aggregation		182	17	3.9×10^{-9}
Response to elevated platelet cytosolic Ca ²⁺		82	9	5.5×10^{-8}
Platelet homeostasis		53	7	5.7×10^{-8}
Nucleotide-like (purinergic) receptors		16	4	1.8×10^{-7}
Platelet degranulation		77	8	2.8×10^{-7}
Peptide ligand-binding receptors		175	14	3.8×10^{-7}
Molecules associated with elastic fibres		34	5	7.1×10^{-7}
Amine ligand-binding receptors		35	5	8.6×10^{-7}
G _{αi} signalling events		184	14	9.8×10^{-7}
GPCR ligand binding		363	27	1.1×10^{-6}
Elastic fibre formation		38	5	1.5×10^{-6}
G _{αq} signalling events		159	12	1.9×10^{-6}
Serotonin receptors		12	3	3.8×10^{-6}
P2Y receptors		12	3	3.8×10^{-6}
Signal amplification		16	3	2.3×10^{-5}
Gastrin-CREB signalling pathway via PKC and MAPK		180	12	2.3×10^{-5}
Complement cascade		33	4	2.4×10^{-5}
Glycosaminoglycan metabolism		110	8	2.5×10^{-5}
Glycogen breakdown (glycogenolysis)		17	3	2.7×10^{-5}

E.2.1 Resampling Analysis

Table E.5: Pathways for *CDH1* partners from SLIPT in stomach cancer

Reactome Pathway	Over-representation	Permutation
<i>Extracellular matrix organization</i>	7.5×10^{-140}	0.070215
Hemostasis	1.8×10^{-121}	0.25804
Developmental Biology	9.2×10^{-107}	0.53032
Axon guidance	1.5×10^{-102}	0.6704
Eukaryotic Translation Termination	1.9×10^{-99}	$> 1.031 \times 10^{-5}$
GPCR ligand binding	3.8×10^{-99}	0.54914
Viral mRNA Translation	3.3×10^{-98}	$> 1.031 \times 10^{-5}$
Formation of a pool of free 40S subunits	3.3×10^{-98}	$> 1.031 \times 10^{-5}$
Eukaryotic Translation Elongation	1.6×10^{-97}	$> 1.031 \times 10^{-5}$
Peptide chain elongation	7.2×10^{-97}	$> 1.031 \times 10^{-5}$
Class A/1 (Rhodopsin-like receptors)	2.7×10^{-96}	0.58174
Nonsense Mediated Decay independent of the Exon Junction Complex	3×10^{-96}	$> 1.031 \times 10^{-5}$
Infectious disease	2.6×10^{-94}	0.25484
GTP hydrolysis and joining of the 60S ribosomal subunit	3.4×10^{-94}	$> 1.031 \times 10^{-5}$
L13a-mediated translational silencing of Ceruloplasmin expression	2.8×10^{-92}	$> 1.031 \times 10^{-5}$
3' -UTR-mediated translational regulation	2.8×10^{-92}	$> 1.031 \times 10^{-5}$
Neuronal System	8.4×10^{-92}	0.53433
SRP-dependent cotranslational protein targeting to membrane	9.5×10^{-92}	$> 1.031 \times 10^{-5}$
Eukaryotic Translation Initiation	2.0×10^{-90}	$> 1.031 \times 10^{-5}$
Cap-dependent Translation Initiation	2.0×10^{-90}	$> 1.031 \times 10^{-5}$
Nonsense-Mediated Decay	7.4×10^{-90}	$> 1.031 \times 10^{-5}$
Nonsense Mediated Decay enhanced by the Exon Junction Complex	7.4×10^{-90}	$> 1.031 \times 10^{-5}$
Adaptive Immune System	8.1×10^{-88}	0.14116
Translation	1.3×10^{-87}	$> 1.031 \times 10^{-5}$
Platelet activation, signalling and aggregation	1.3×10^{-86}	0.28959
Influenza Infection	1×10^{-82}	$> 1.031 \times 10^{-5}$
Influenza Viral RNA Transcription and Replication	2.4×10^{-82}	$> 1.031 \times 10^{-5}$
Influenza Life Cycle	2×10^{-80}	$> 1.031 \times 10^{-5}$
Response to elevated platelet cytosolic Ca ²⁺	4.9×10^{-78}	0.50817
Signalling by NGF	1.6×10^{-75}	0.38518
Rho GTPase cycle	5.1×10^{-75}	0.14864
Signalling by PDGF	7.4×10^{-74}	0.40493
<i>Signalling by Rho GTPases</i>	5.1×10^{-73}	0.077217
Glycosaminoglycan metabolism	1.4×10^{-68}	0.52984
<i>G_{ai} signalling events</i>	1.8×10^{-66}	0.9254
Metabolism of carbohydrates	1.1×10^{-65}	0.39501
G_{as} signalling events	2.7×10^{-65}	0.0050293
Potassium Channels	2.7×10^{-65}	0.53359
Transmission across Chemical Synapses	1.8×10^{-64}	0.81833
ECM proteoglycans	3.4×10^{-64}	0.083482
Peptide ligand-binding receptors	4.8×10^{-64}	0.62817
Degradation of the extracellular matrix	1.1×10^{-63}	0.80879
Platelet homeostasis	5.3×10^{-63}	0.53134
NGF signalling via TRKA from the plasma membrane	6.1×10^{-63}	0.5717
Integration of energy metabolism	4.5×10^{-61}	0.10889
Collagen formation	5.4×10^{-61}	0.29896
Integrin cell surface interactions	7×10^{-59}	0.18167
Collagen biosynthesis and modifying enzymes	7×10^{-59}	0.30208
Neurotransmitter Receptor Binding And Downstream Transmission	8.7×10^{-57}	0.82522
In The Postsynaptic Cell		
Signalling by Wnt	8.7×10^{-57}	0.25468

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Table E.6: Pathways for *CDH1* partners from SLIPT in stomach and siRNA screen

Reactome Pathway	Over-representation	Permutation
Platelet activation, signalling and aggregation	3.9×10^{-9}	0.49557
Class A/1 (Rhodopsin-like receptors)	3.9×10^{-9}	0.98432
Response to elevated platelet cytosolic Ca ²⁺	5.5×10^{-8}	0.54349
Platelet homeostasis	5.7×10^{-8}	0.45017
Nucleotide-like (purinergic) receptors	1.8×10^{-7}	0.36966
Peptide ligand-binding receptors	3.8×10^{-7}	0.91294
Molecules associated with elastic fibres	7.1×10^{-7}	0.0025868
Amine ligand-binding receptors	8.6×10^{-7}	0.43303
G _{ai} signalling events	9.8×10^{-7}	0.99626
GPCR ligand binding	1.1×10^{-6}	0.97733
Elastic fibre formation	1.5×10^{-6}	0.0025868
G _{aq} signalling events	1.9×10^{-6}	0.86089
P2Y receptors	3.8×10^{-6}	0.18795
Serotonin receptors	3.8×10^{-6}	0.37853
Signal amplification	2.3×10^{-5}	0.47856
Gastrin-CREB signalling pathway via PKC and MAPK	2.3×10^{-5}	0.98567
Complement cascade	2.4×10^{-5}	$> 3.4628 \times 10^{-6}$
Glycosaminoglycan metabolism	2.5×10^{-5}	0.38953
Glycogen breakdown (glycogenolysis)	2.7×10^{-5}	0.83772
Defective B4GALT7 causes EDS, progeroid type	4.9×10^{-5}	0.10792
Defective B3GAT3 causes JDSSDHD	4.9×10^{-5}	0.10792
Role of LAT2/NTAL/LAB on calcium mobilization	5.6×10^{-5}	0.35373
Cell surface interactions at the vascular wall	5.6×10^{-5}	0.47642
G_{as} signalling events	6×10^{-5}	0.019858
Signalling by NOTCH	6×10^{-5}	0.19008
A tetrasaccharide linker sequence is required for GAG synthesis	0.00017	0.47642
Extracellular matrix organization	0.00018	0.0047308
Collagen formation	0.00018	0.19245
Effects of PIP2 hydrolysis	0.0002	0.37779
Syndecan interactions	0.0002	0.37779
Diseases associated with glycosaminoglycan metabolism	0.00023	0.01028
Diseases of glycosylation	0.00023	0.01028
<i>Chondroitin sulfate/dermatan sulfate metabolism</i>	0.00023	0.085541
Integrin alphaIIb beta3 signalling	0.00028	0.76936
Keratan sulfate biosynthesis	0.00034	0.68744
Rho GTPase cycle	0.00034	0.15675
Creation of C4 and C2 activators	0.00035	0.12275
Abacavir transport and metabolism	0.00035	0.12443
Amine compound SLC transporters	0.00037	0.69773
FCER1 mediated NF-κB activation	0.00037	0.69846
Fc epsilon receptor (FCER1) signalling	0.00056	0.43303
Defective EXT2 causes exostoses 2	0.00067	0.16053
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	0.00067	0.16053
<i>Collagen biosynthesis and modifying enzymes</i>	0.00071	0.052911
Keratan sulfate/keratin metabolism	0.00073	0.46533
G alpha (12/13) signalling events	0.00078	0.59164
SEMA3A-Plexin repulsion signalling by inhibiting Integrin adhesion	0.00084	0.038504
Signal attenuation	0.00084	0.37779
Eicosanoid ligand-binding receptors	0.0011	0.11117
SOS-mediated signalling	0.0011	0.25387

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italicics (FDR < 0.1).

E.3 Metagene Analysis

Metagene analysis was also performed for synthetic lethal candidates for *CDH1* expression in stomach cancer.

Table E.7: Candidate synthetic lethal metagenes against *CDH1* from SLIPT in stomach cancer

Pathway	ID	Observed	Expected	χ^2 value	p-value	p-value (FDR)
Cell-Cell communication	1500931	18	50.4	110	7.43×10^{-23}	1.53×10^{-20}
VEGFR2 mediated vascular permeability	5218920	19	50.4	109	1.36×10^{-22}	2.49×10^{-20}
Sema4D in semaphorin signalling	400685	20	50.4	104	1.62×10^{-21}	2.12×10^{-19}
Ion transport by P-type ATPases	936837	17	50.4	100	8.29×10^{-21}	8.06×10^{-19}
Sialic acid metabolism	4085001	19	50.4	95.3	9.95×10^{-20}	7.82×10^{-18}
Synthesis of pyrophosphates in the cytosol	1855167	26	50.4	94	1.86×10^{-19}	1.23×10^{-17}
Keratan sulfate/keratin metabolism	1638074	25	50.4	93.5	2.36×10^{-19}	1.44×10^{-17}
Ion channel transport	983712	19	50.4	92.8	3.37×10^{-19}	1.99×10^{-17}
Keratan sulfate biosynthesis	2022854	26	50.4	91.4	6.79×10^{-19}	3.62×10^{-17}
Arachidonic acid metabolism	2142753	22	50.4	90.6	9.81×10^{-19}	5.07×10^{-17}
RHO GTPases activate CIT	5625900	22	50.4	87	5.80×10^{-18}	2.66×10^{-16}
Stimuli-sensing channels	2672351	25	50.4	85.8	1.03×10^{-17}	4.58×10^{-16}
Synthesis of PI	1483226	19	50.4	85.6	1.15×10^{-17}	4.89×10^{-16}
G-protein activation	202040	19	50.4	85.3	1.34×10^{-17}	5.53×10^{-16}
NrCAM interactions	447038	22	50.4	84.3	2.1×10^{-17}	8.27×10^{-16}
Inwardly rectifying K^+ channels	1296065	24	50.4	83.5	3.19×10^{-17}	1.22×10^{-15}
Calcitonin-like ligand receptors	419812	20	50.4	82.2	6.07×10^{-17}	2.13×10^{-15}
Prostacyclin signalling through prostacyclin receptor	392851	24	50.4	81.8	7.27×10^{-17}	2.5×10^{-15}
Presynaptic function of Kainate receptors	500657	26	50.4	79.7	2.00×10^{-16}	6.34×10^{-15}
ADP signalling through P2Y purinoceptor 12	392170	23	50.4	79.2	2.57×10^{-16}	7.71×10^{-15}
regulation of FZD by ubiquitination	4641263	22	50.4	78.8	3.15×10^{-16}	9.3×10^{-15}
Toxicity of tetanus toxin (TeNT)	5250982	27	50.4	78.7	3.36×10^{-16}	9.75×10^{-15}
Gap junction degradation	190873	21	50.4	78.5	3.66×10^{-16}	1.04×10^{-14}
Nephrin interactions	373753	25	50.4	78.2	4.21×10^{-16}	1.14×10^{-14}
GABA synthesis, release, reuptake and degradation	888590	26	50.4	77	7.69×10^{-16}	1.95×10^{-14}

Strongest candidate SL partners for *CDH1* by SLIPT with observed and expected numbers of TCGA stomach cancer samples with low expression of both genes.

Appendix F

Synthetic Lethal Genes in Pathways

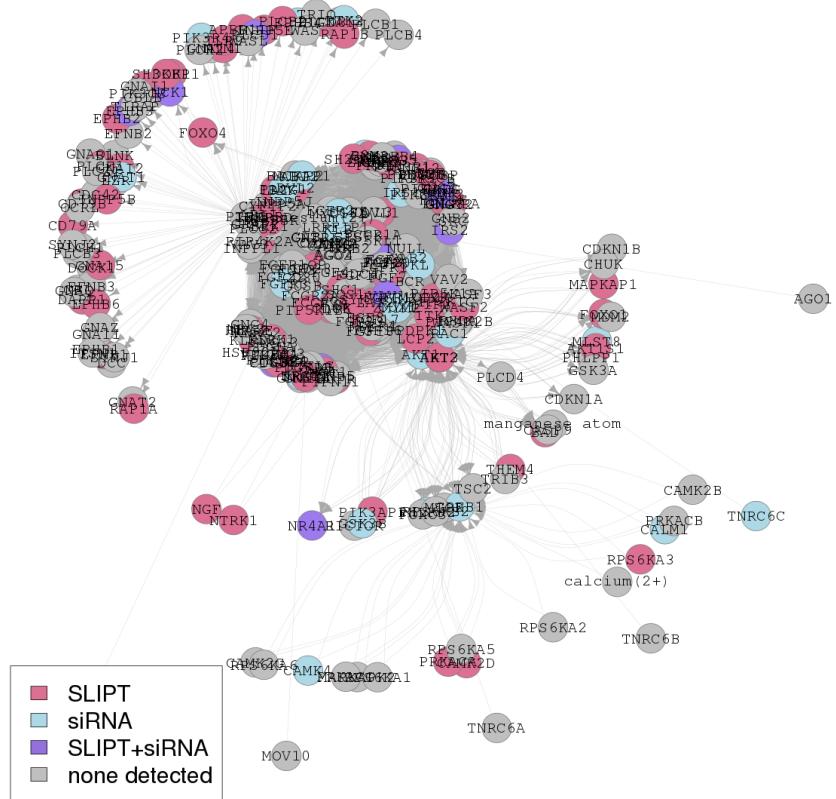


Figure F.1: Synthetic Lethality in the PI3K/AKT Pathway. The Reactome PI3K/AKT pathway with synthetic lethal candidates coloured as shown in the legend.

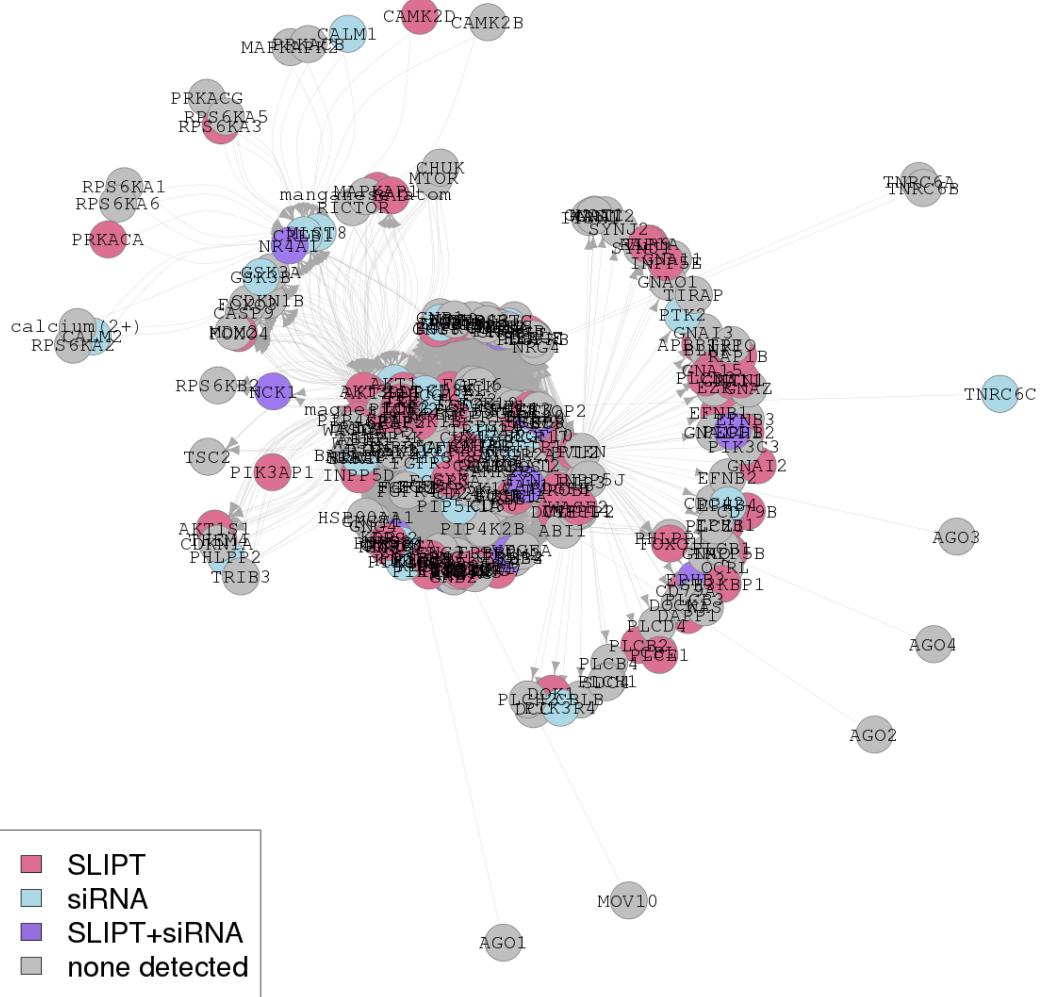


Figure F.2: Synthetic Lethality in the PI3K/AKT Pathway in Cancer. The Reactome PI3K/AKT Pathway in Cancer pathway with synthetic lethal candidates coloured as shown in the legend.

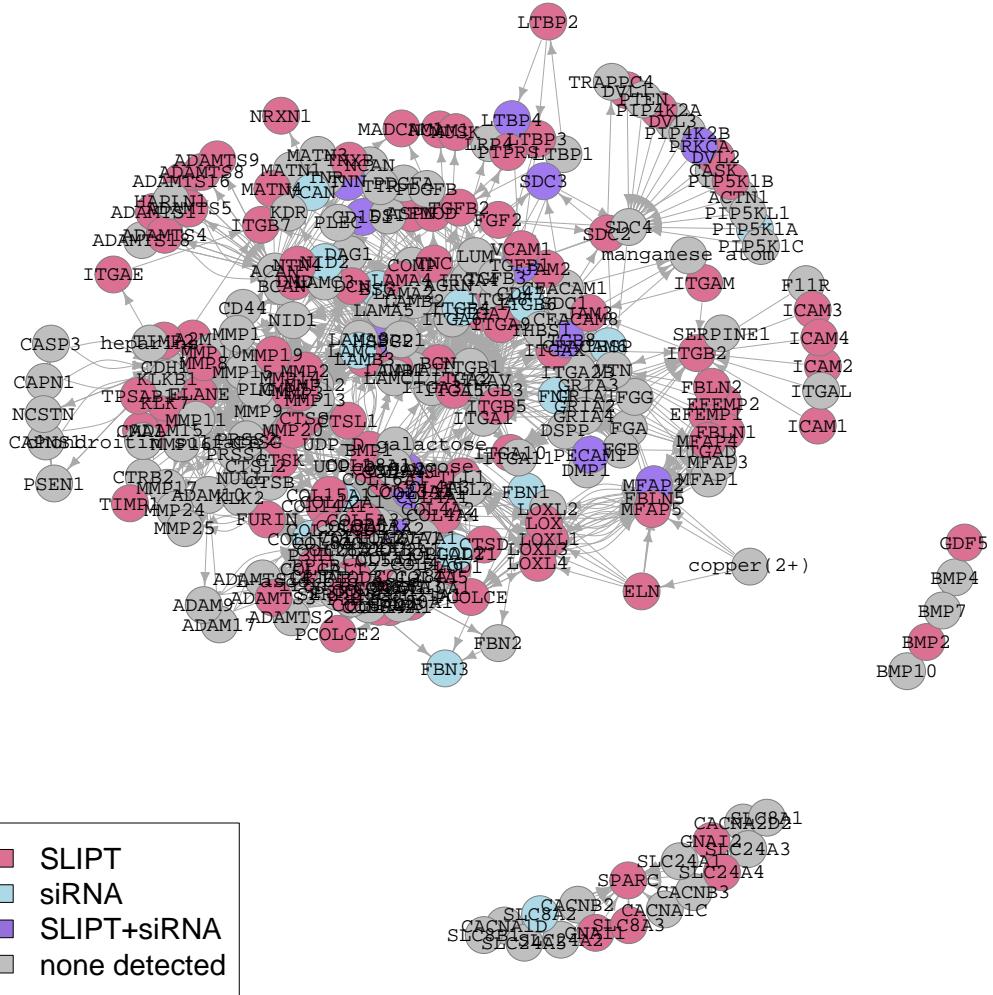


Figure F.3: Synthetic Lethality in the Extracellular Matrix. The Reactome Extracellular Matrix pathway with synthetic lethal candidates coloured as shown in the legend.

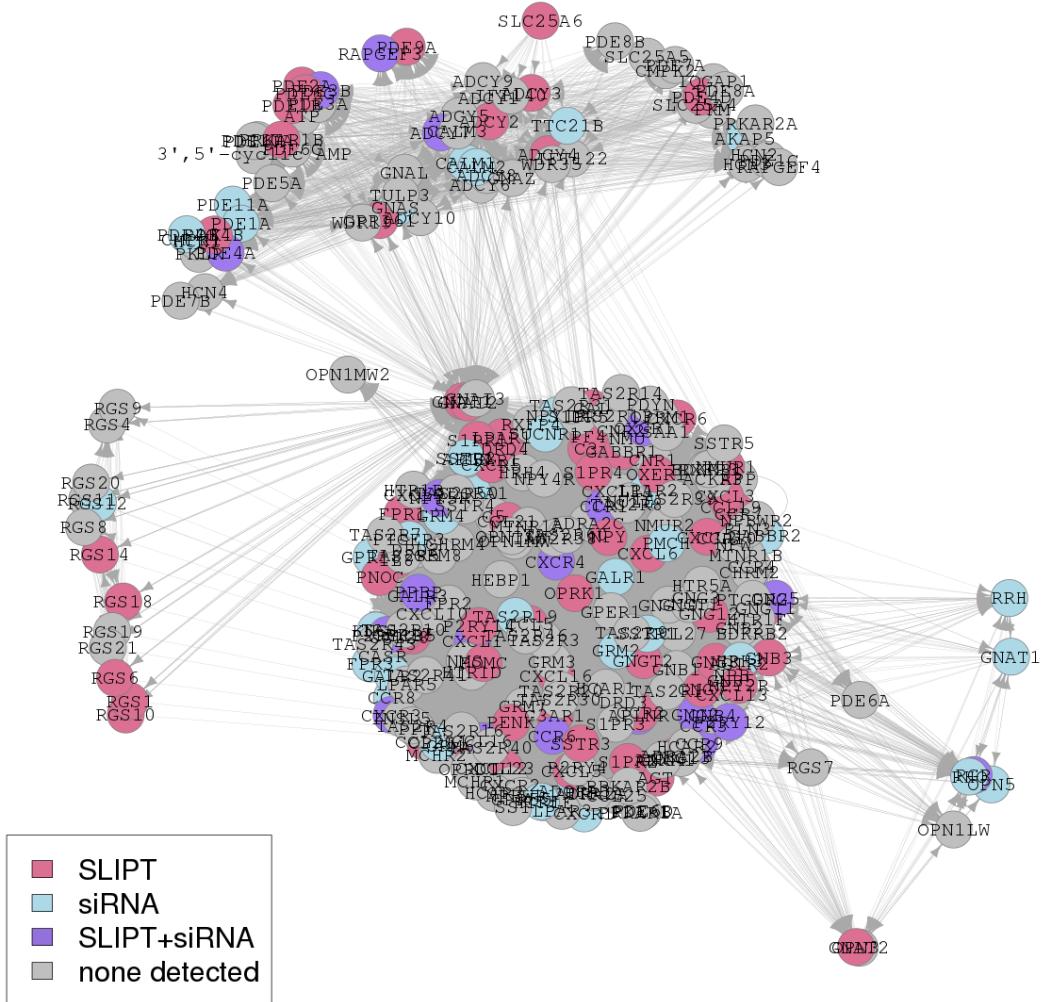


Figure F.4: Synthetic Lethality in the GPCRs. The Reactome $G_{\alpha i}$ pathway with synthetic lethal candidates coloured as shown in the legend.

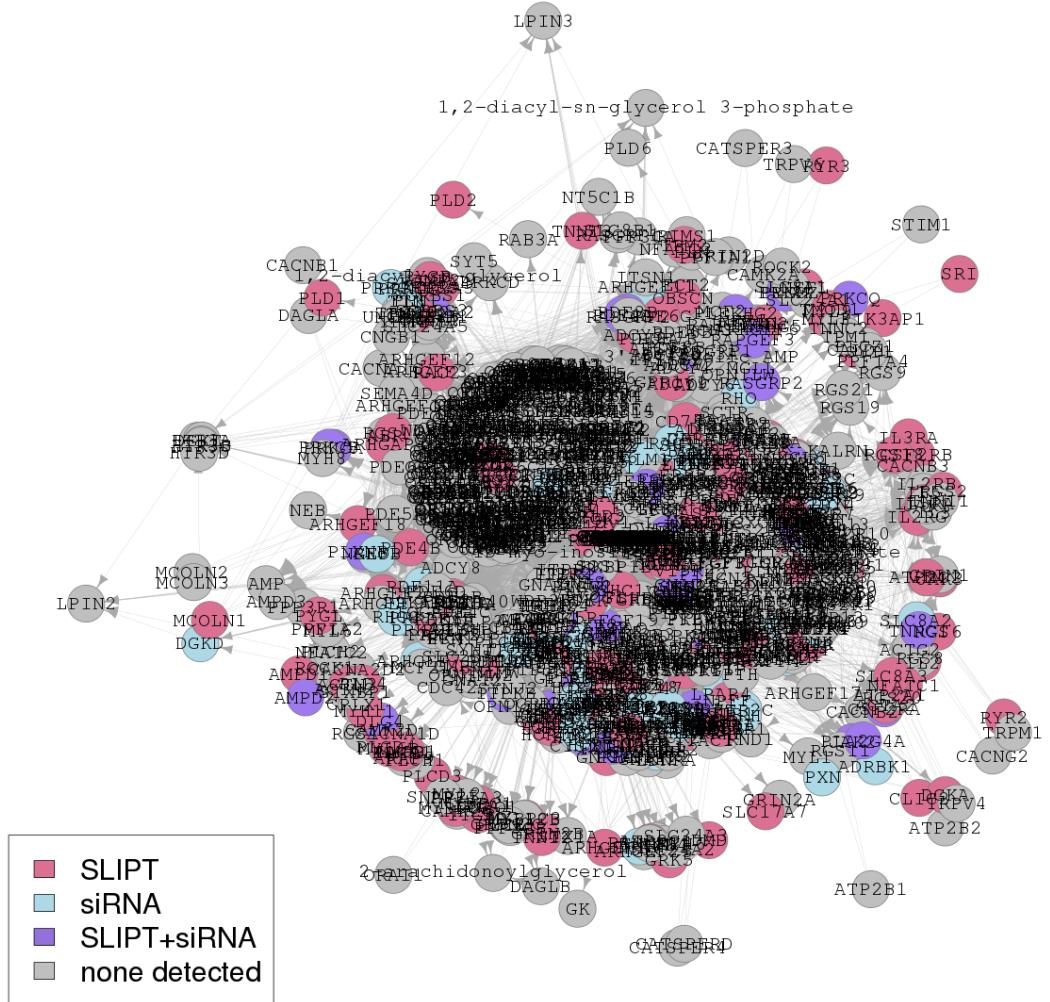


Figure F.5: **Synthetic Lethality in the GPCR Downstream.** The Reactome GPCR Downstream pathway with synthetic lethal candidates coloured as shown in the legend.

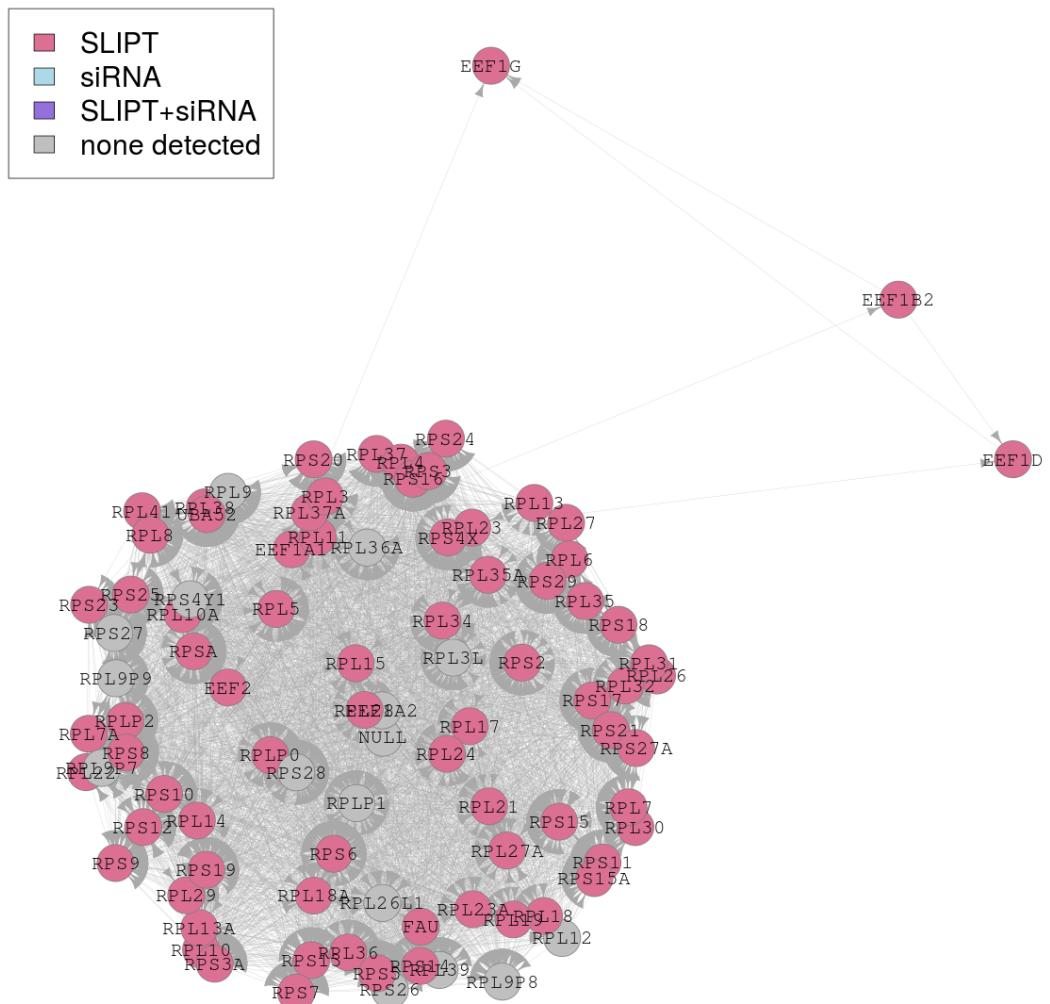


Figure F.6: **Synthetic Lethality in the Translation Elongation**. The Reactome Translation Elongation pathway with synthetic lethal candidates coloured as shown in the legend.

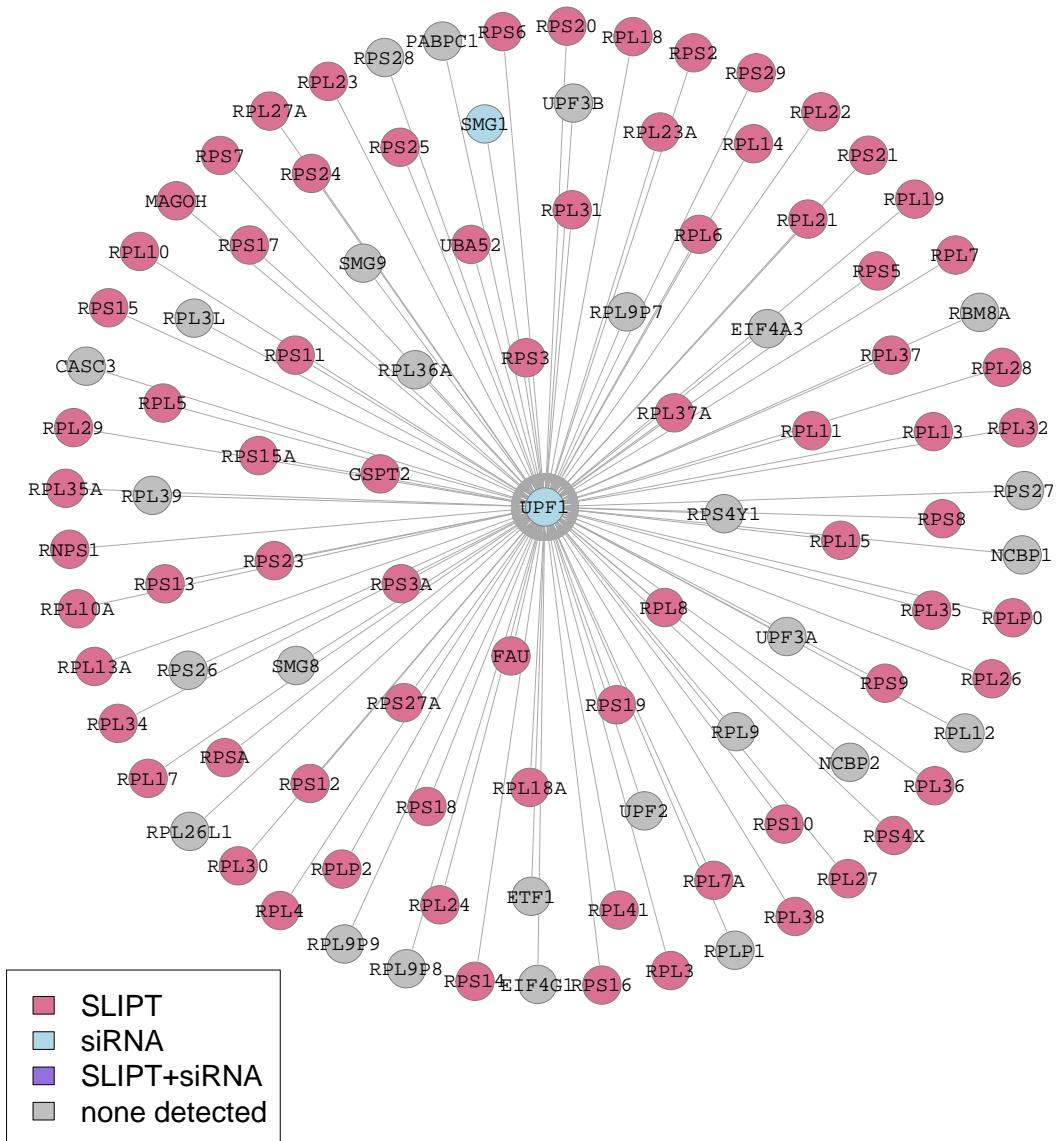


Figure F.7: Synthetic Lethality in the Nonsense-mediated Decay. The Reactome Nonsense-mediated Decay pathway with synthetic lethal candidates coloured as shown in the legend.

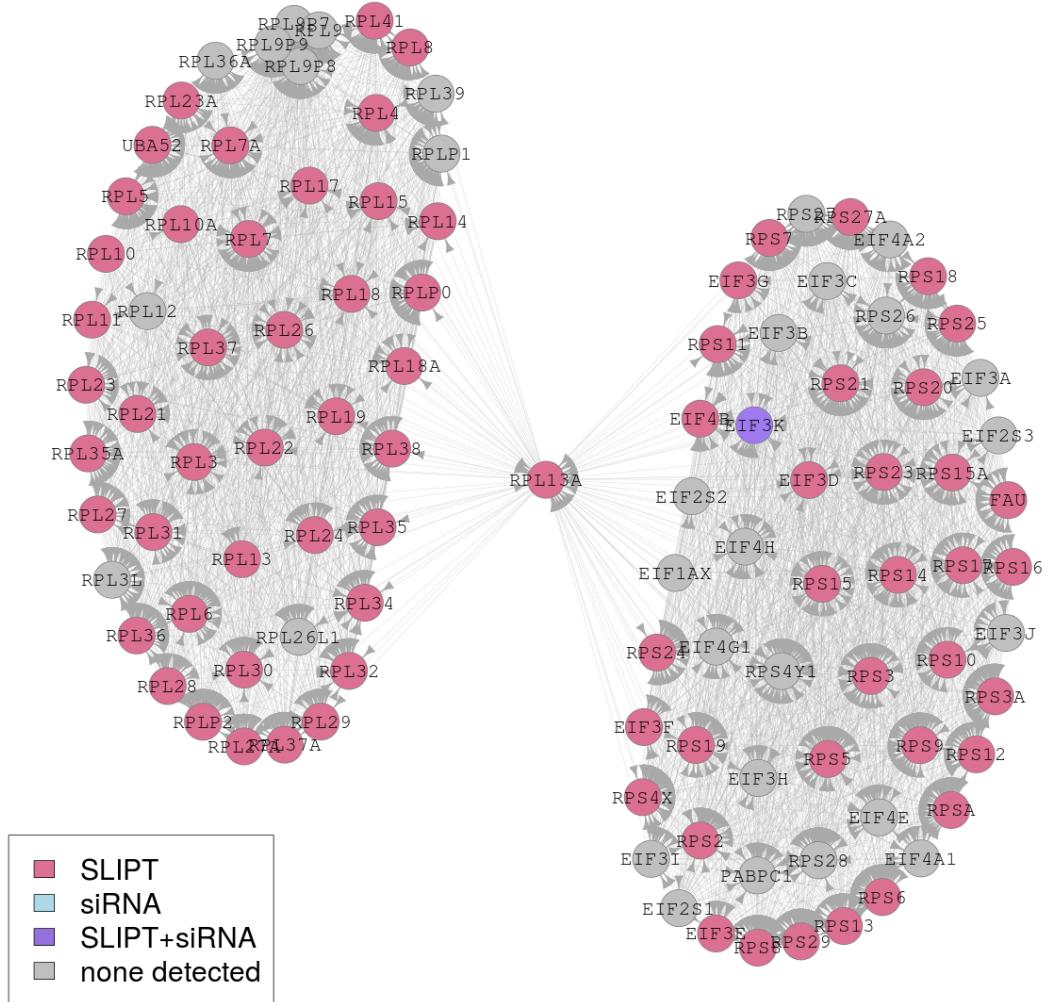


Figure F.8: **Synthetic Lethality in the 3' UTR.** The Reactome 3' UTR pathway with synthetic lethal candidates coloured as shown in the legend.

Appendix G

Pathway Connectivity for Mutation SLIPT

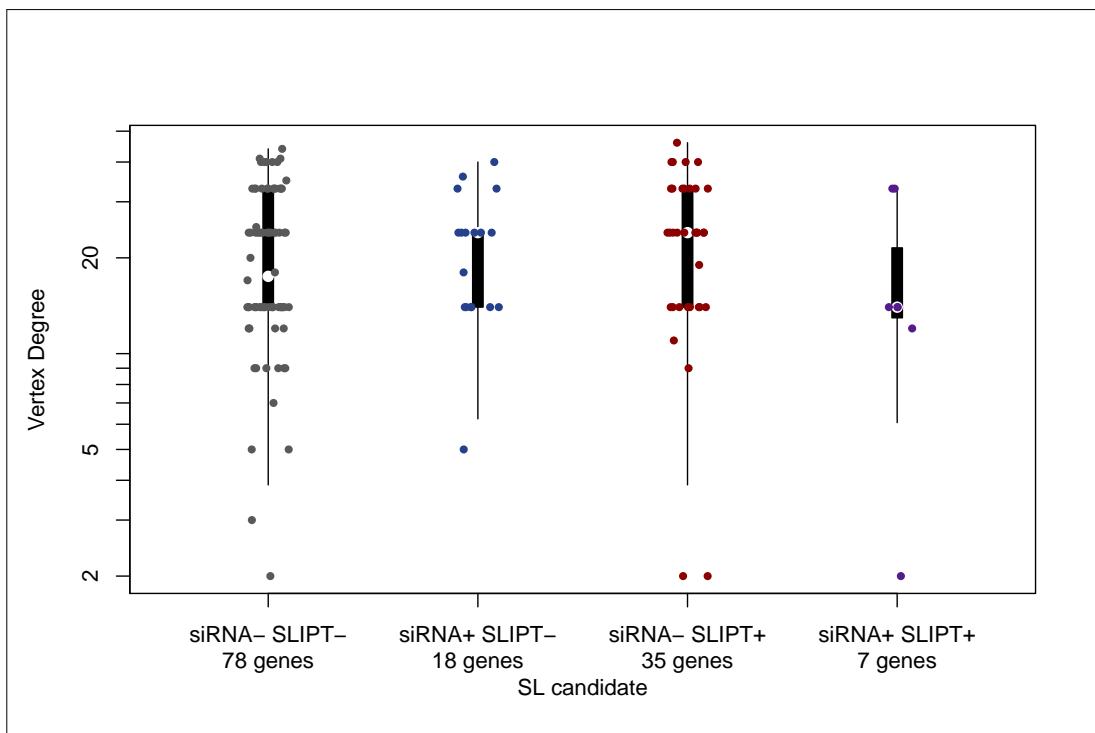


Figure G.1: **Synthetic Lethality and Vertex Degree.** The number of connected genes (vertex degree) was compared (on a log-scale across genes detected by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. There were very few differences in vertex degree between the groups, although genes detected by siRNA included those with the fewest connections.

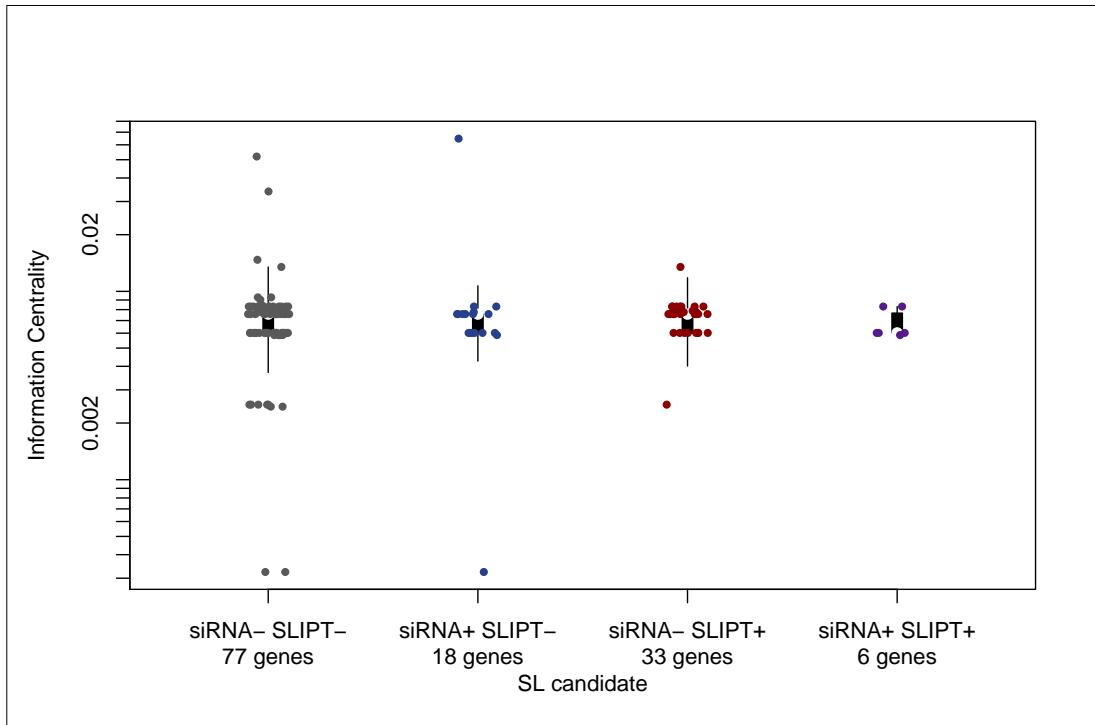


Figure G.2: Synthetic Lethality and Centrality. The information centrality was compared (on a log-scale across genes detected by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by mtSLIPT or siRNA did not have higher connectivity than genes not detected by either approach. The gene with the highest centrality was detected by siRNA.

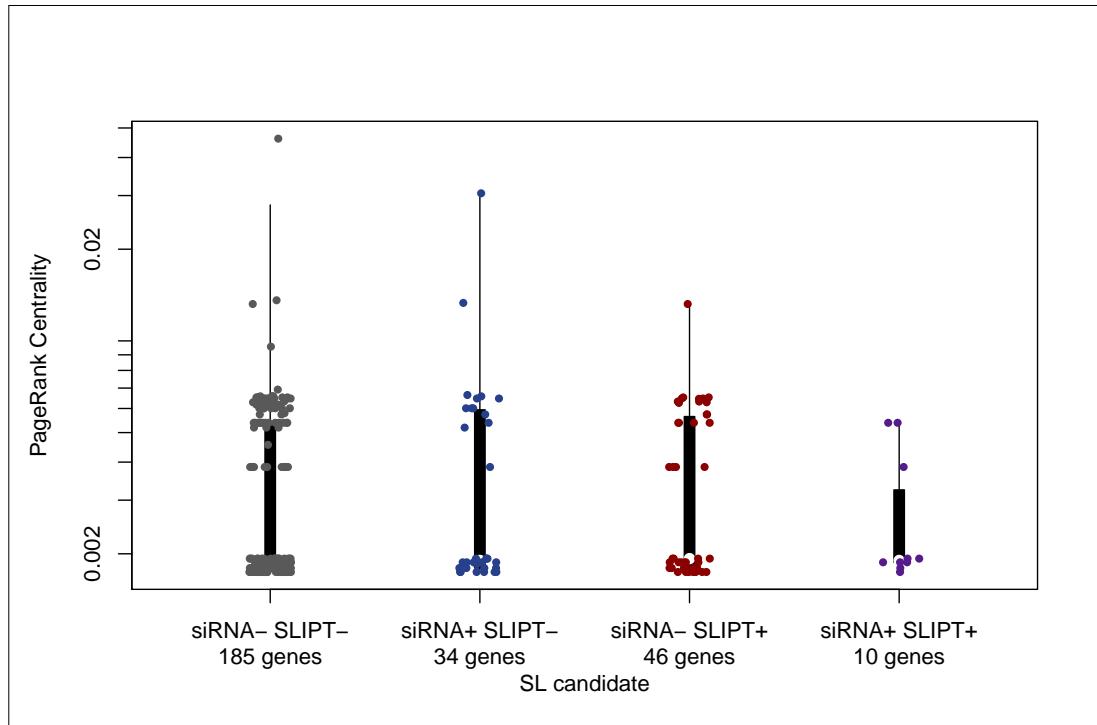


Figure G.3: **Synthetic Lethality and PageRank.** The PageRank centrality was compared (on a log-scale across genes detected by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had a more restricted range of centrality values than other genes not detected by either approach, although these groups also had fewer genes.

Table G.1: ANOVA for Synthetic Lethality and Vertex Degree

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	15	15.50	0.0134	0.9084
mtSLIPT	1	196	195.94	0.1689	0.6825
siRNA×mtSLIPT	1	9	9.17	0.0079	0.9294

Analysis of variance for vertex degree against synthetic lethal detection approaches (with an interaction term)

Table G.2: ANOVA for Synthetic Lethality and Information Centrality

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.000256	0.0002561	0.1851	0.6685
mtSLIPT	1	0.003225	0.0032247	2.3308	0.1318
siRNA×mtSLIPT	1	0.001238	0.0012385	0.8952	0.3476

Analysis of variance for information centrality against synthetic lethal detection approaches (with an interaction term)

Table G.3: ANOVA for Synthetic Lethality and PageRank Centrality

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.0002038	2.0385×10^{-4}	1.1423	0.2892
mtSLIPT	1	0.0000208	2.0752×10^{-5}	0.1163	0.7342
siRNA×mtSLIPT	1	0.0000137	1.3743×10^{-5}	0.0770	0.7823

Analysis of variance for PageRank centrality against synthetic lethal detection approaches (with an interaction term)

Appendix H

Information Centrality for Gene Essentiality

Network structure is another useful strategy to analyse gene function and this has been used to investigate network properties of a network constructed from of Reactome pathways imported via Pathway Commons with Paxtools (Cerami *et al.*, 2011; Demir *et al.*, 2013). Most notably, information centrality which has been proposed as a measure of gene essentiality was calculated as performed by Kranthi *et al.* (2013) using the efficiency and shortest path between each pair of nodes in the network before and after a node of interest is removed to test the importance of a node to network connectivity. Reactome contains substrates and cofactors in addition to genes or proteins. In support of centrality as a measure of essentiality, a number nodes with the highest centrality (shown in Table H.1) were essential nutrients including Mg^{2+} , Ca^{2+} , Zn^{2+} , and Fe. In addition, there were genes important in development of epithelial tissues and breast cancer such as *IL8*, *GATA3*, and *CTNNB1* detected with relatively high information centrality.

Table H.1: Information centrality for genes and molecules in the Reactome network

Node	Centrality
<i>ZNF473</i>	0.0510
magnesium(2+)	0.0082
<i>XBP1</i>	0.0053
calcium(2+)	0.0050
zinc(2+)	0.0048
iron atom	0.0041
<i>FMN</i>	0.0040
<i>AGT</i>	0.0037
<i>HSP90AA1</i>	0.0029
phosphatidyl-L-serine	0.0029
<i>P2RX7</i>	0.0026
<i>PANX1</i>	0.0024
<i>NCAM1</i>	0.0022
<i>NUDT1</i>	0.0021
<i>PLAUR</i>	0.0020
<i>IL8</i>	0.0020
<i>HSPA8</i>	0.0019
<i>TYROBP</i>	0.0019
<i>CASP3</i>	0.0017
<i>GNAL</i>	0.0015
<i>CBLB</i>	0.0015
<i>HBB</i>	0.0014
<i>GATA4</i>	0.0013
<i>TGS1</i>	0.0013
<i>CTNNB1</i>	0.0012

Highest information centrality for genes (proteins), cofactors, and minerals in the Reactome network

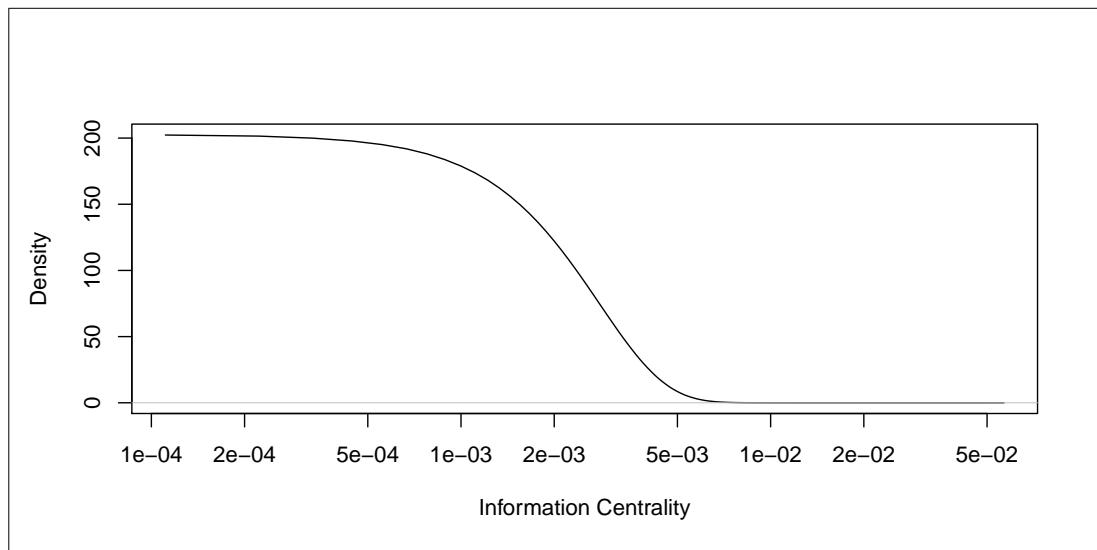


Figure H.1: **Information centrality distribution.** Information centrality in the Reactome network for nodes, including genes/proteins and other biomolecules.

Appendix I

Pathway Structure for Mutation SLIPT

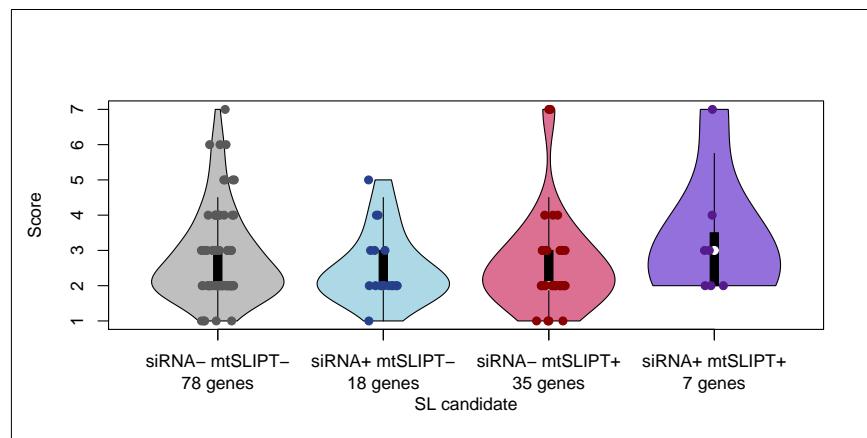


Figure I.1: **Synthetic Lethality and Heirarchy Score in PI3K.** The hierarchical distance scores were similarly distributed across mtSLIPT and siRNA genes. Genes detected by both methods had a higher (downstream) median than either group.

Table I.1: ANOVA for Synthetic Lethality and PI3K Hierarchy

	DF	Sum Squares	Mean Squares	F-value	p-value
siRNA	1	0.001	0.00070	0.0004	0.9841
mtSLIPT	1	0.007	0.0066	0.0040	0.9496
siRNA×mtSLIPT	1	3.906	3.9056	2.3829	0.1250

Analysis of variance for PI3K hierarchy score against synthetic lethal detection approaches (with an interaction term)

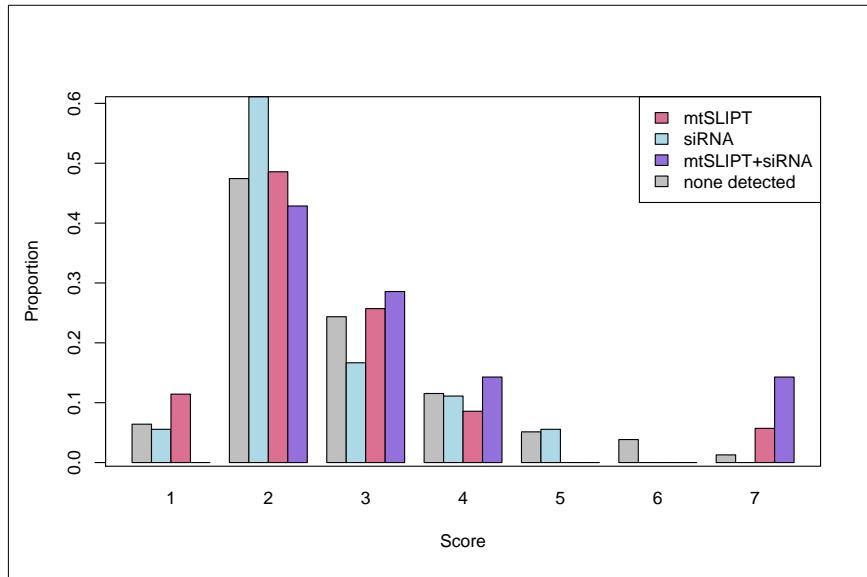


Figure I.2: **Heirarchy Score in PI3K against Synthetic Lethality in PI3K.** The number of mtSLIPT and siRNA genes against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

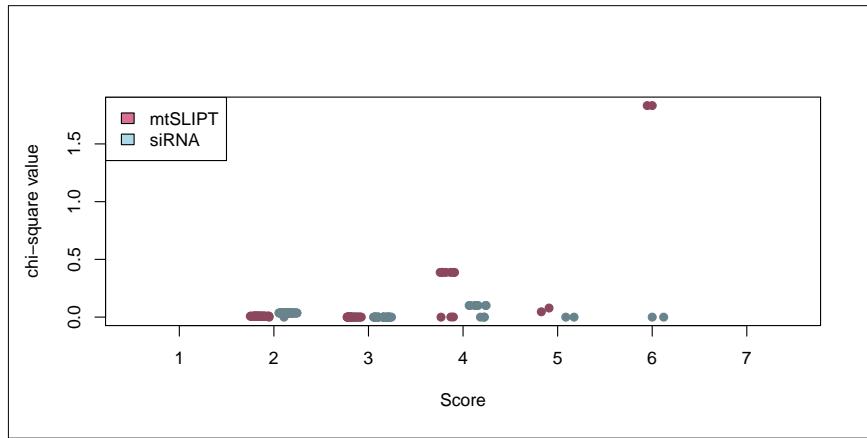


Figure I.3: **Structure of Synthetic Lethality in PI3K.** The number of mtSLIPT and siRNA genes against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities. The number of mtSLIPT and siRNA genes upstream or downstream of each gene in the Reactome PI3K pathway were tested (by the χ^2 -test). These are plotted as a split jitter stripchart against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

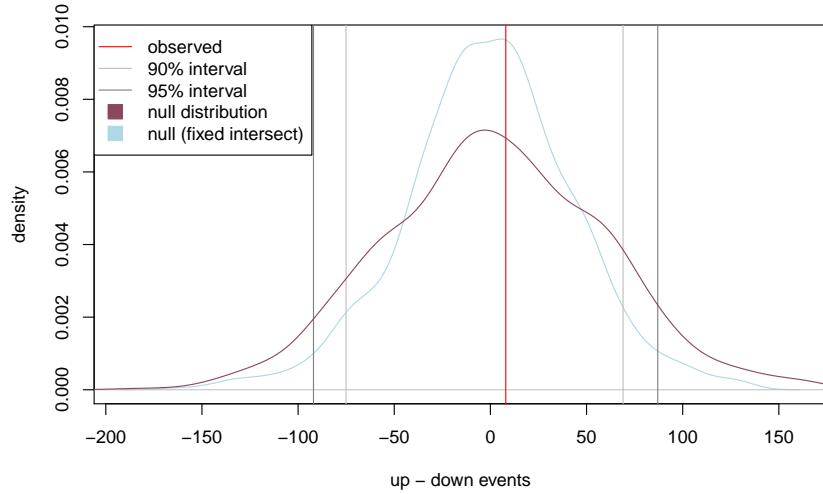


Figure I.4: Structure of Synthetic Lethality Resampling. A null distribution (10,000 iterations) of the siRNA genes upstream or downstream of mtSLIPT genes (shown by the difference) in the PI3K pathway. The observed events (red) were compared to the the distribution (violet) and were not significant. Genes detected by both methods were fixed for the distribution (blue). The genes detected by both approaches were used.

Table I.2: Resampling for pathway structure of synthetic lethal detection methods

Pathway	Graph		States		Observed				Permutation p-value	
	Nodes	Edges	mtSL	siRNA	Up	Down	Up-Down	Up/Down	Up-Down	Down-Up
PI3K Cascade	138	1495	42	25	131	123	8	1.065	0.4473	0.5466
PI3K/AKT Signalling in Cancer	275	12882	56	44	478	440	38	1.086	0.4163	0.5810
G_{αi} Signalling	292	22003	57	58	543	866	-323	0.627	0.9507	0.0488
GPCR downstream	1270	142071	218	160	7632	6500	1132	1.174	0.1707	0.8291
Elastic fibre formation	42	175	16	7	6	7	-1	0.857	0.5512	0.3681
Extracellular matrix	299	3677	81	29	313	347	-34	0.902	0.5762	0.4215
Formation of Fibrin	52	243	11	5	8	19	-11	0.421	0.7993	0.1800
Nonsense-Mediated Decay	103	102	56	2	0	0	0		0.197	0.1373
3'-UTR-mediated translational regulation	107	2860	56	1	52	1	51	52	0.1210	0.8751
Eukaryotic Translation Elongation	92	3746	57	0	0	0	0		0.4952	0.4892

Pathways in the Reactome network tested for structural relationships between mtSLIPT and siRNA genes by resampling. The raw p-value (computed without adjusting for multiple comparisons over pathways) is given for the difference in upstream and downstream paths from mtSLIPT to siRNA gene candidate partners of CDH1 with significant pathways highlighted in bold. Sampling was performed only in the target pathway and shortest paths were computed within it. Loops or paths in either direction that could not be resolved were excluded from the analysis. The gene detected by both mtSLIPT and siRNA (or resampling for them) were included in the analysis and the number of these were fixed to the number observed.

Appendix J

Performance of SLIPT and χ^2

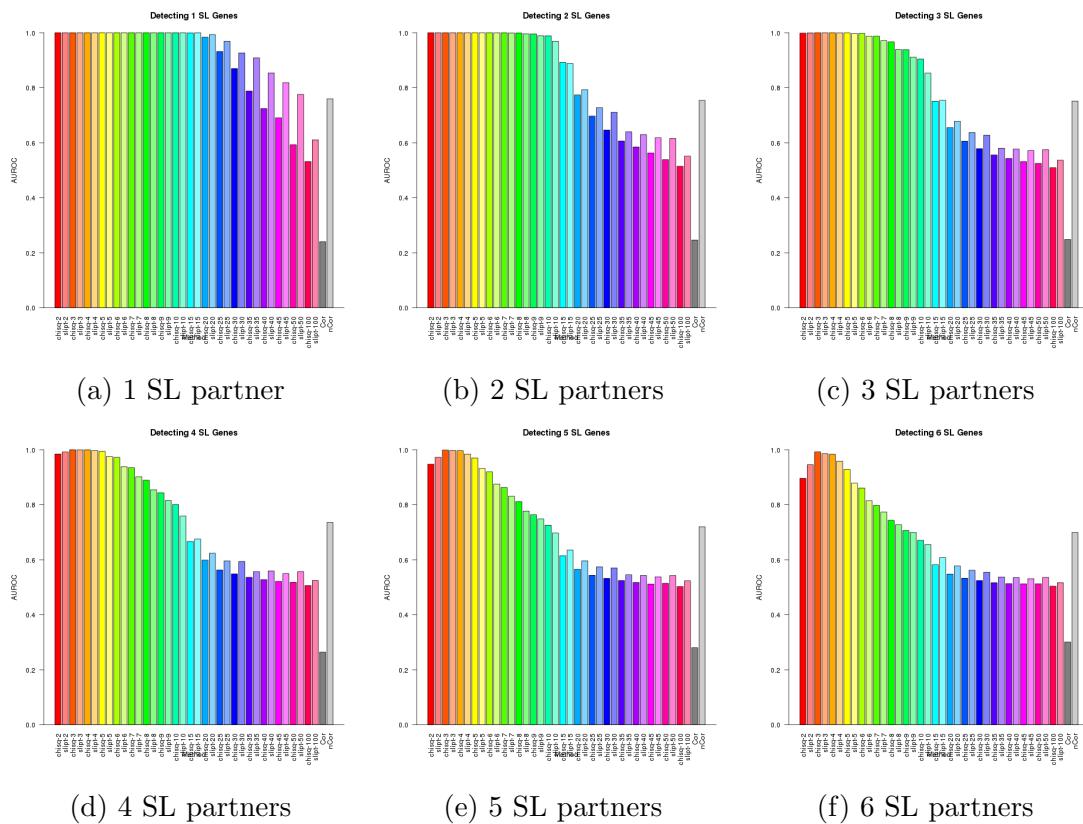


Figure J.1: Performance of χ^2 and SLIPT across quantiles. (continued on next page)

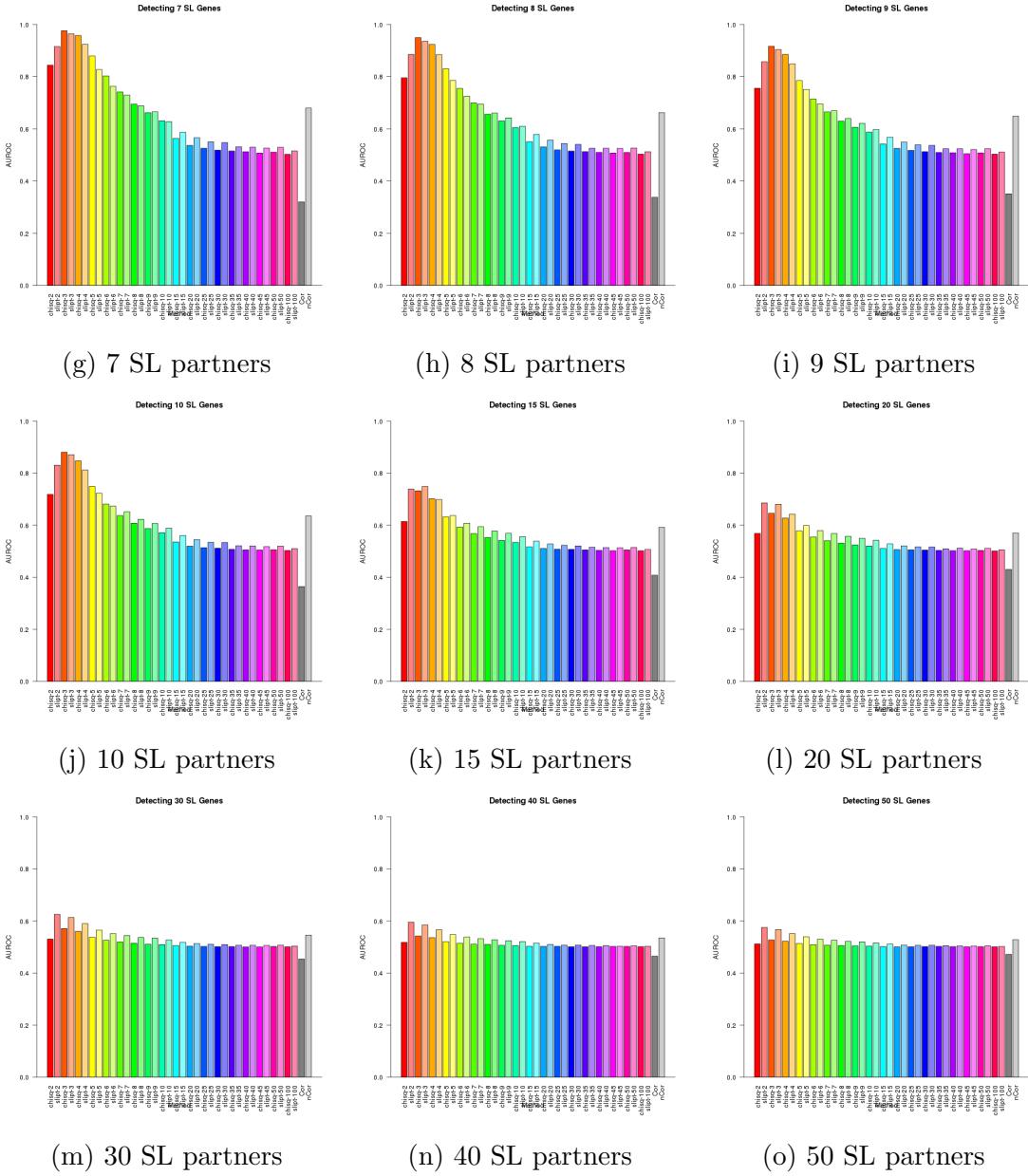


Figure J.1: Performance of χ^2 and SLIPT across quantiles. Synthetic lethal detection with quantiles as in axis labels. The barplot uses the same hues for each quantile (grey for correlation) and darker for χ^2 (and positive correlation). SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 10,000 simulations of 100 genes and 1000 samples. SLIPT performs better than χ^2 for higher numbers of synthetic lethal genes and finer quantiles.

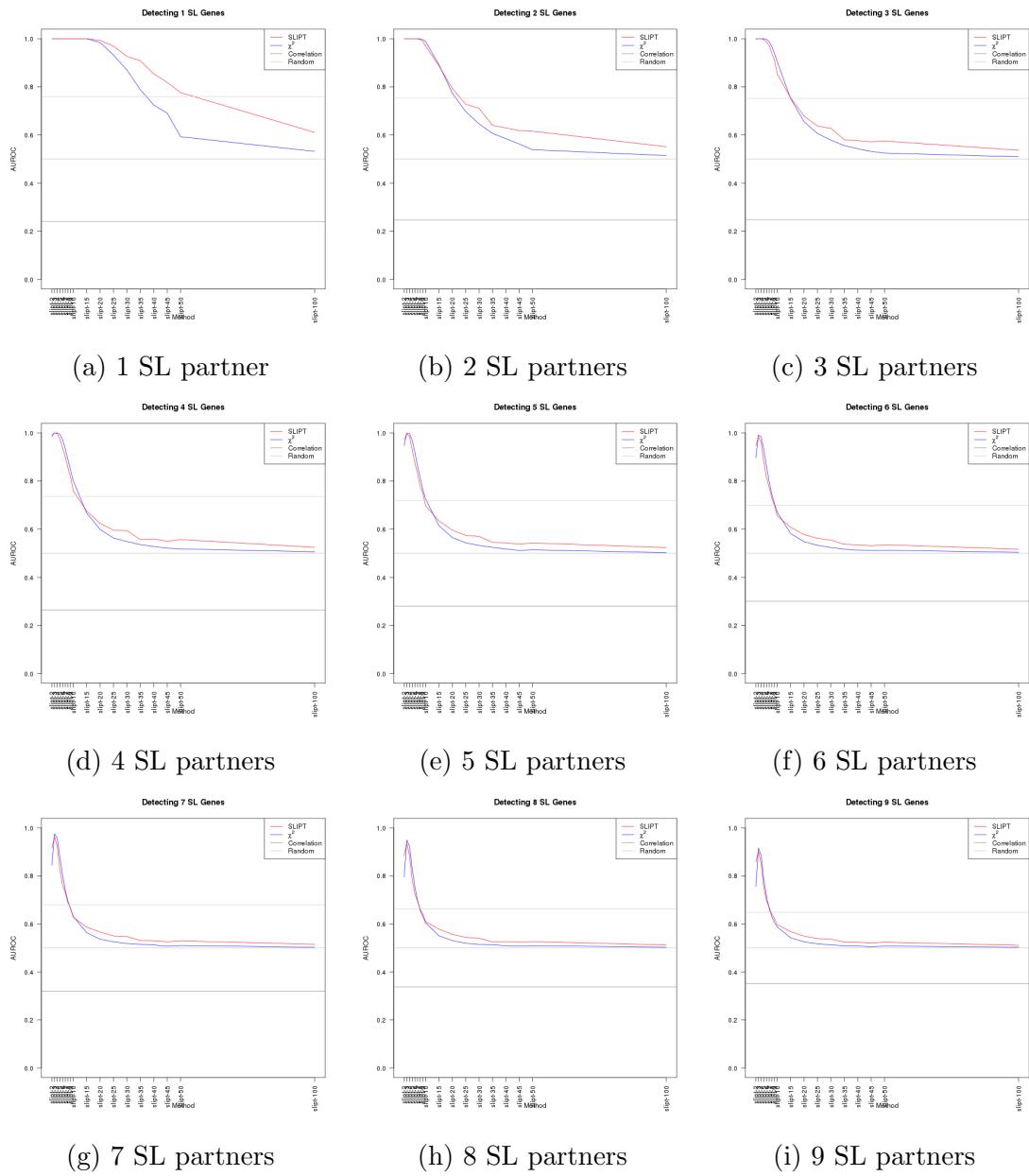


Figure J.2: **Performance of χ^2 and SLIPT across quantiles.** (continued on next page)

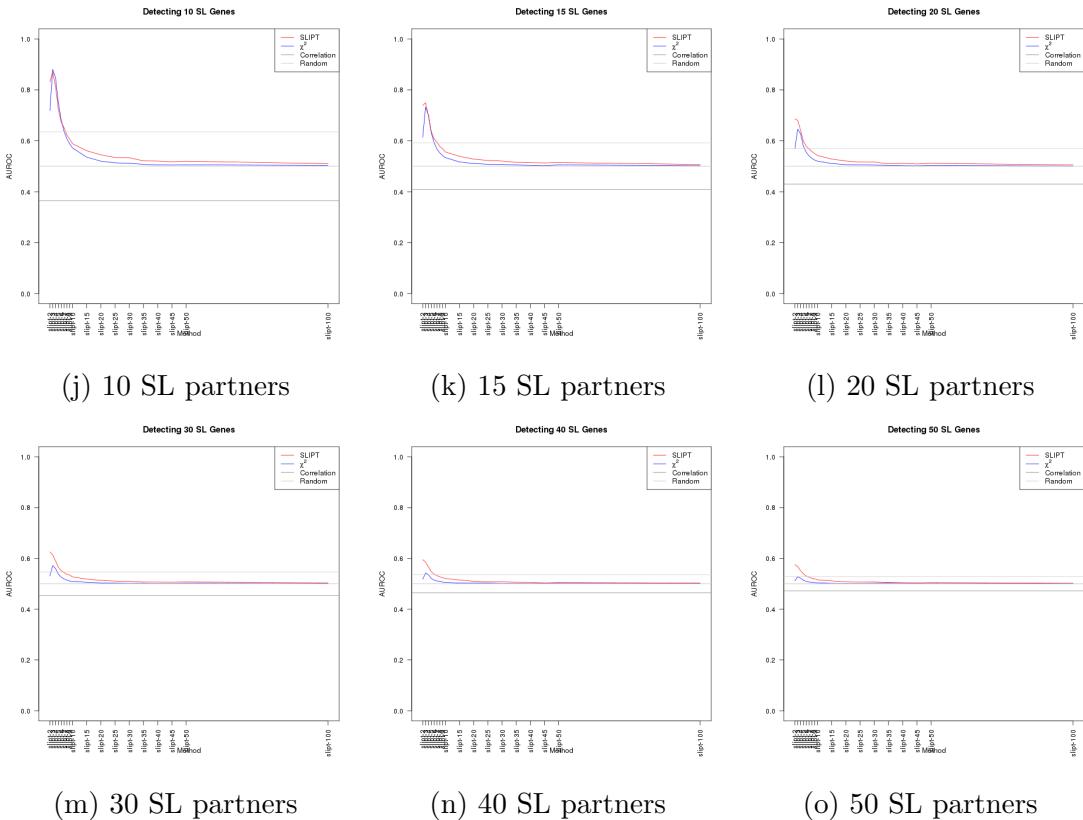


Figure J.2: Performance of χ^2 and SLIPT across quantiles. Synthetic lethal detection with quantiles as in axis labels. The line plots are coloured for SLIPT (red), χ^2 (blue) and correlation (grey) according to the legend. SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 10,000 simulations of 100 genes and 1000 samples. SLIPT performs better than χ^2 for higher numbers of synthetic lethal genes and finer quantiles.

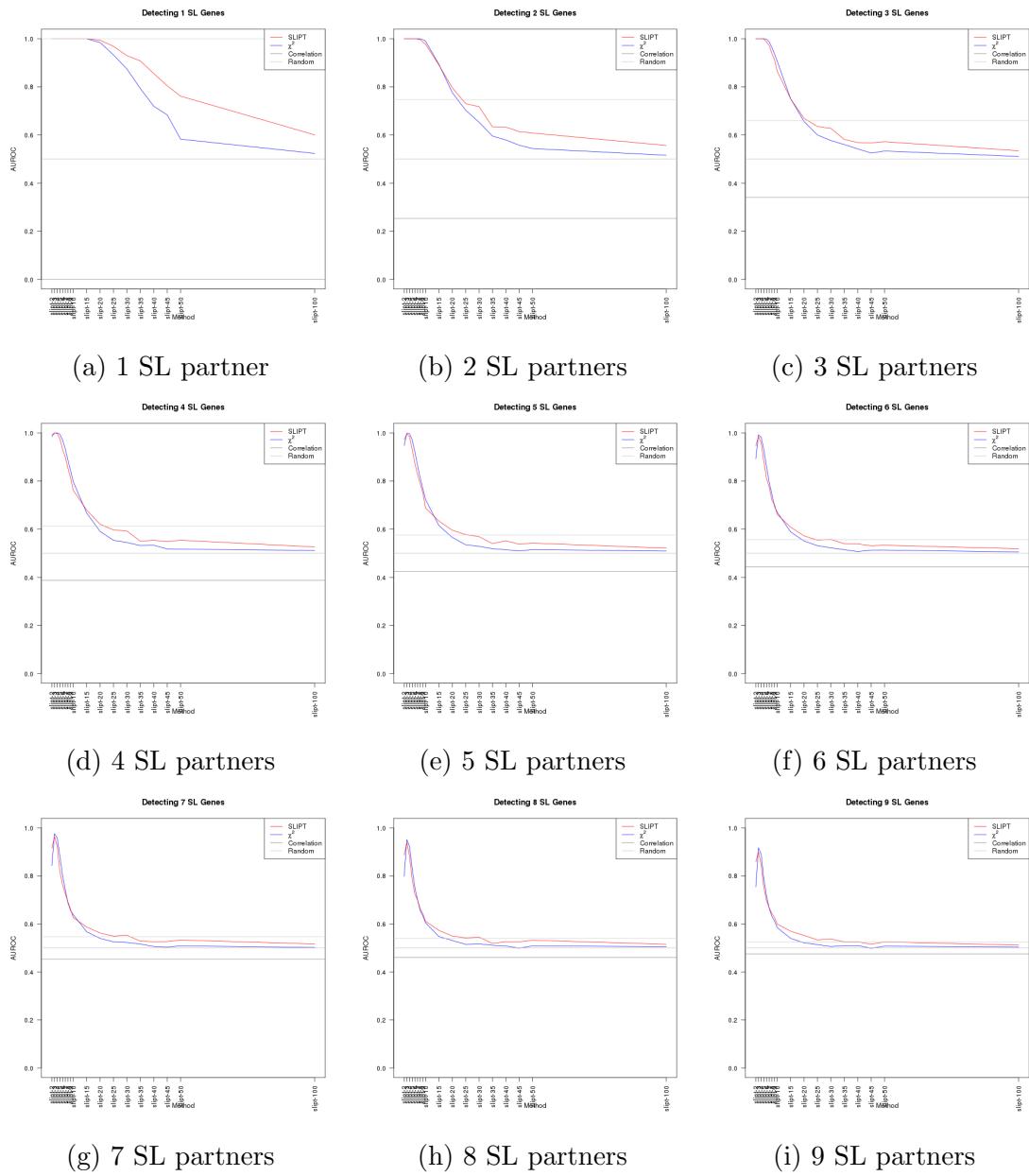


Figure J.3: **Performance of χ^2 and SLIPT across quantiles with more genes.**
 (continued on next page)

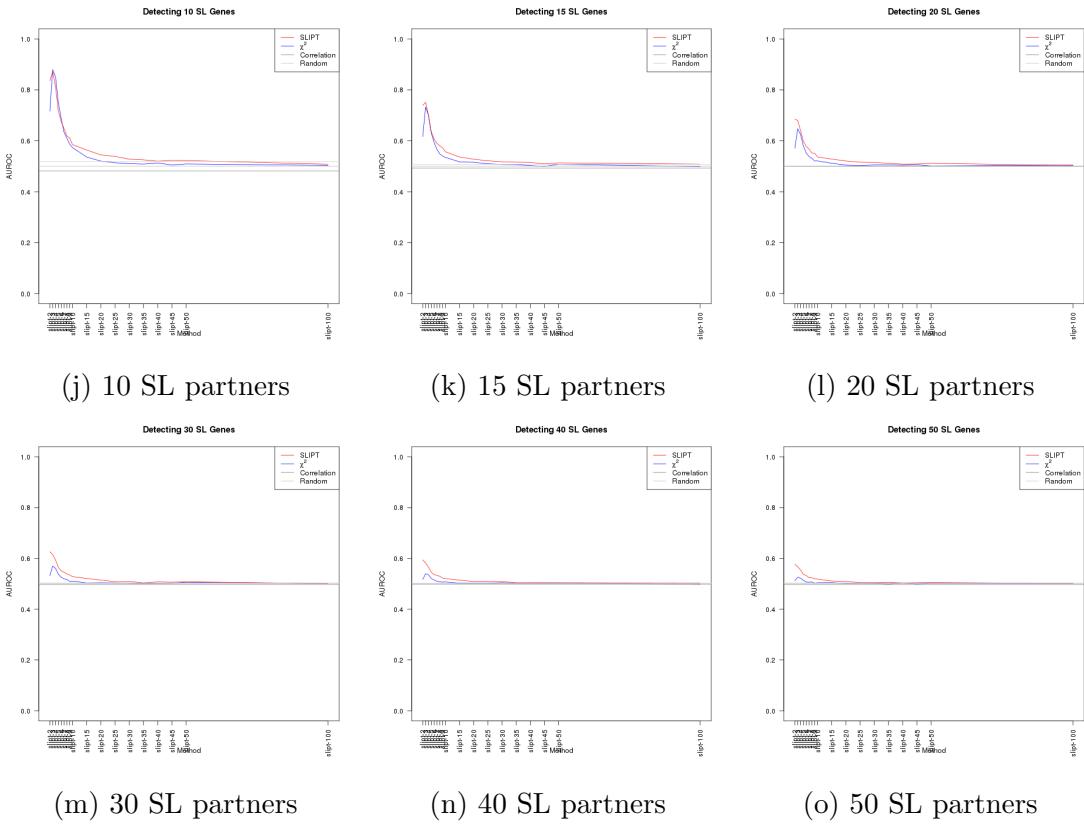


Figure J.3: Performance of χ^2 and SLIPT across quantiles with more genes.
Synthetic lethal detection with quantiles as in axis labels. The line plots are coloured for SLIPT (red), χ^2 (blue) and correlation (grey) according to the legend. SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 1000 simulations of 20,000 genes and 1000 samples. SLIPT performs better than χ^2 for higher numbers of synthetic lethal genes and finer quantiles.

J.1 Correlated Query Genes affects Specificity

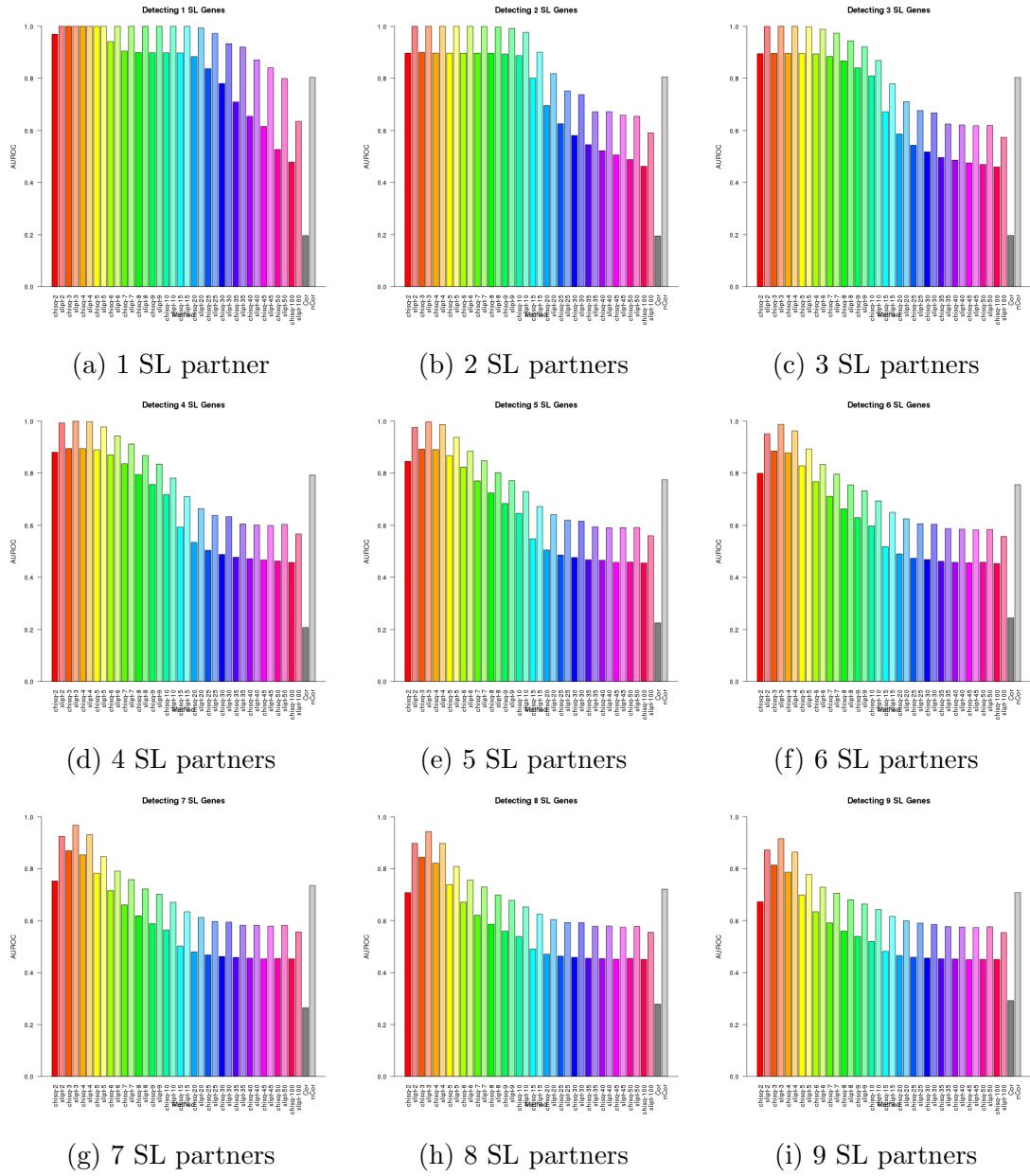


Figure J.4: **Performance of χ^2 and SLIPT across quantiles with query correlation.** (continued on next page)

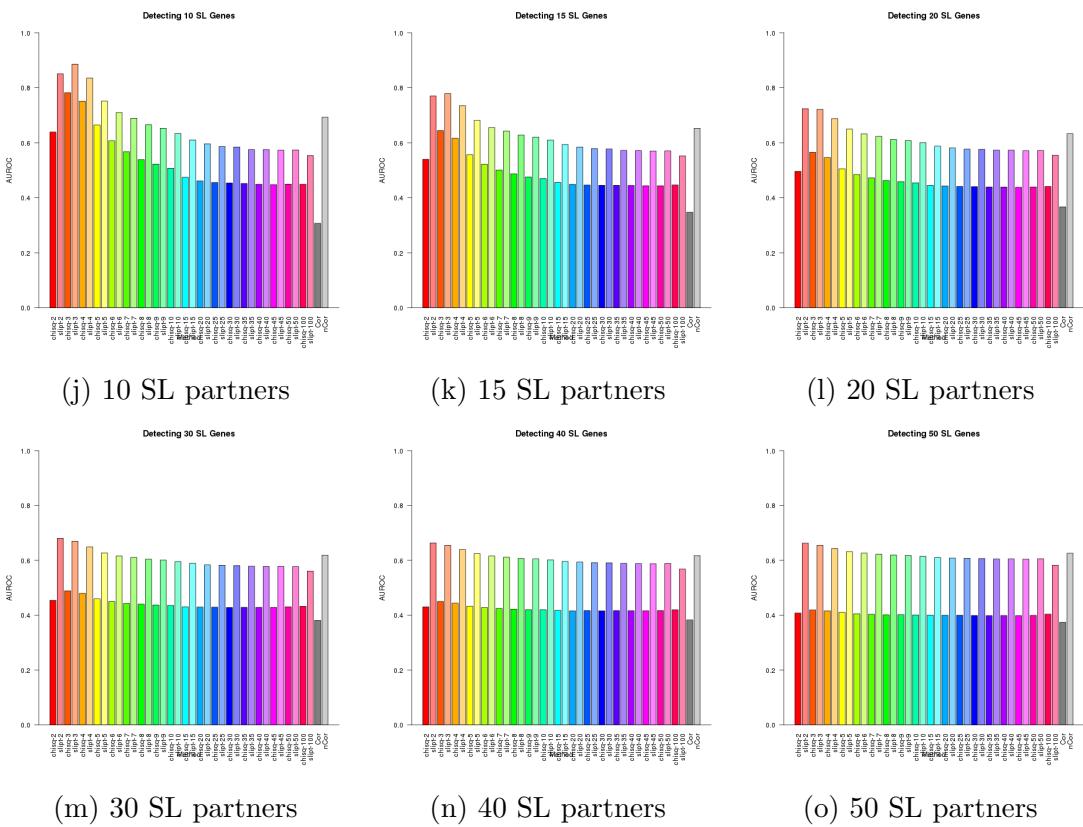


Figure J.4: Performance of χ^2 and SLIPT across quantiles with query correlation. Synthetic lethal detection with quantiles as in axis labels. The barplot uses the same hues for each quantile (grey for correlation) and darker for χ^2 (and positive correlation). SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 10,000 simulations of 100 genes (including 10 correlated with the query) and 1000 samples. SLIPT performs consistently better than χ^2 with positively correlated genes.

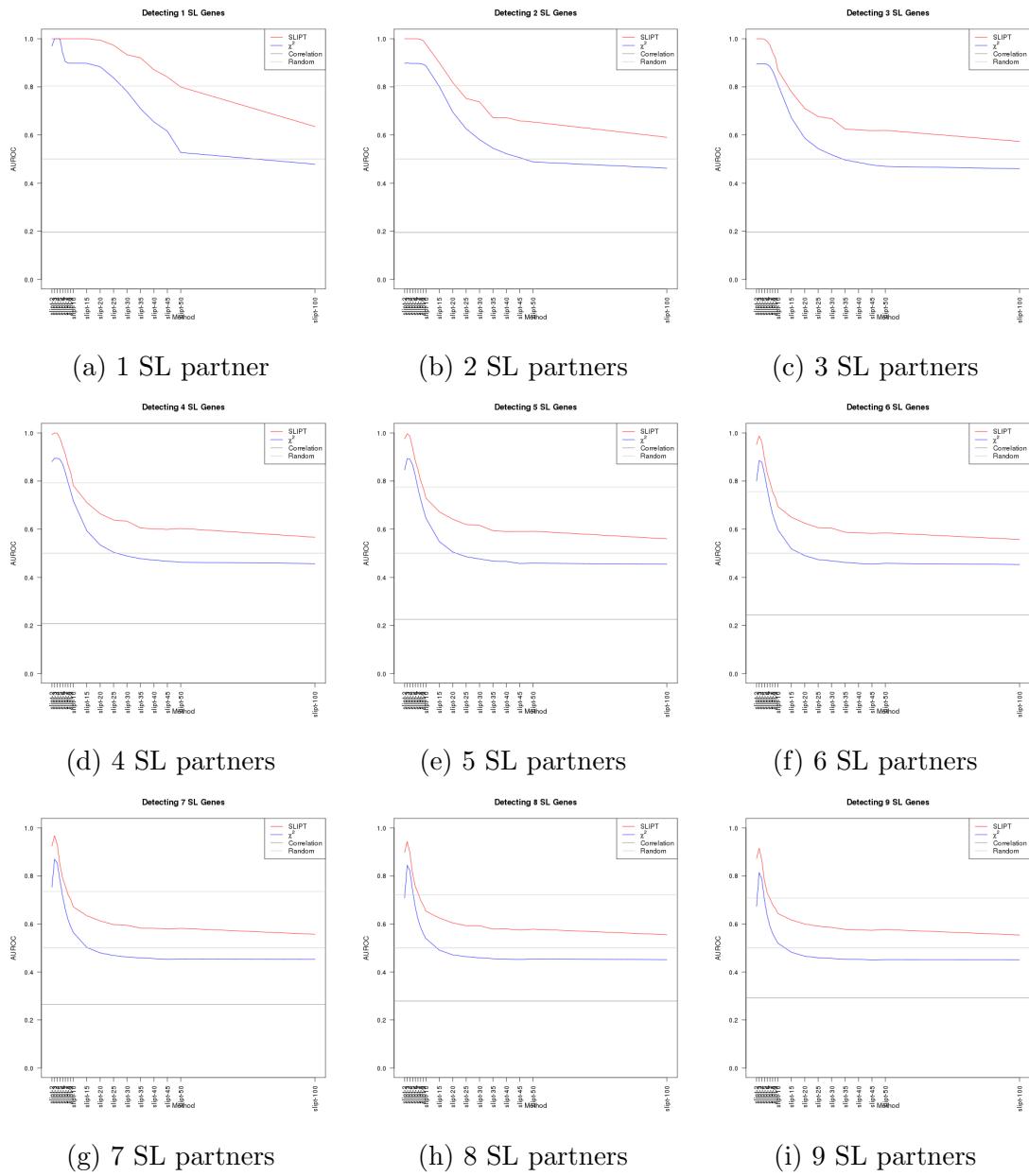


Figure J.5: **Performance of χ^2 and SLIPT across quantiles with query correlation.** (continued on next page)

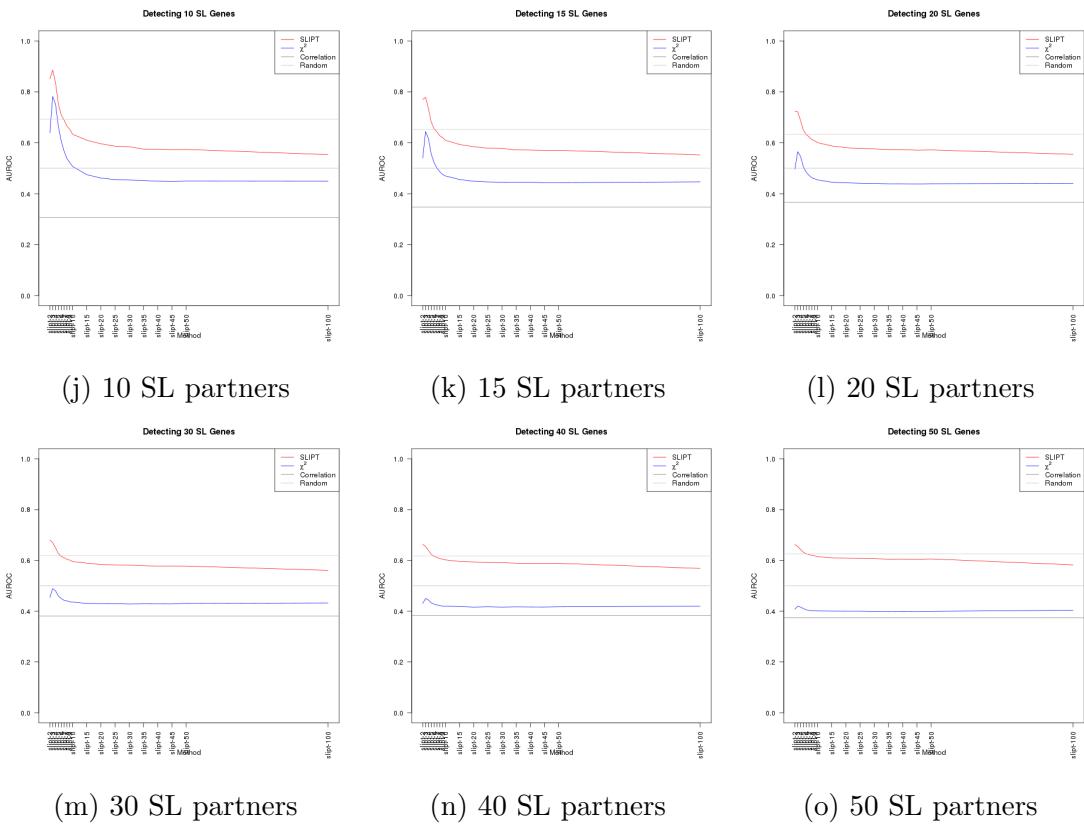


Figure J.5: Performance of χ^2 and SLIPT across quantiles with query correlation. Synthetic lethal detection with quantiles as in axis labels. The line plots are coloured for SLIPT (red), χ^2 (blue) and correlation (grey) according to the legend. SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 10,000 simulations of 100 genes (including 10 correlated with the query) and 1000 samples. SLIPT performs consistently better than χ^2 with positively correlated genes.

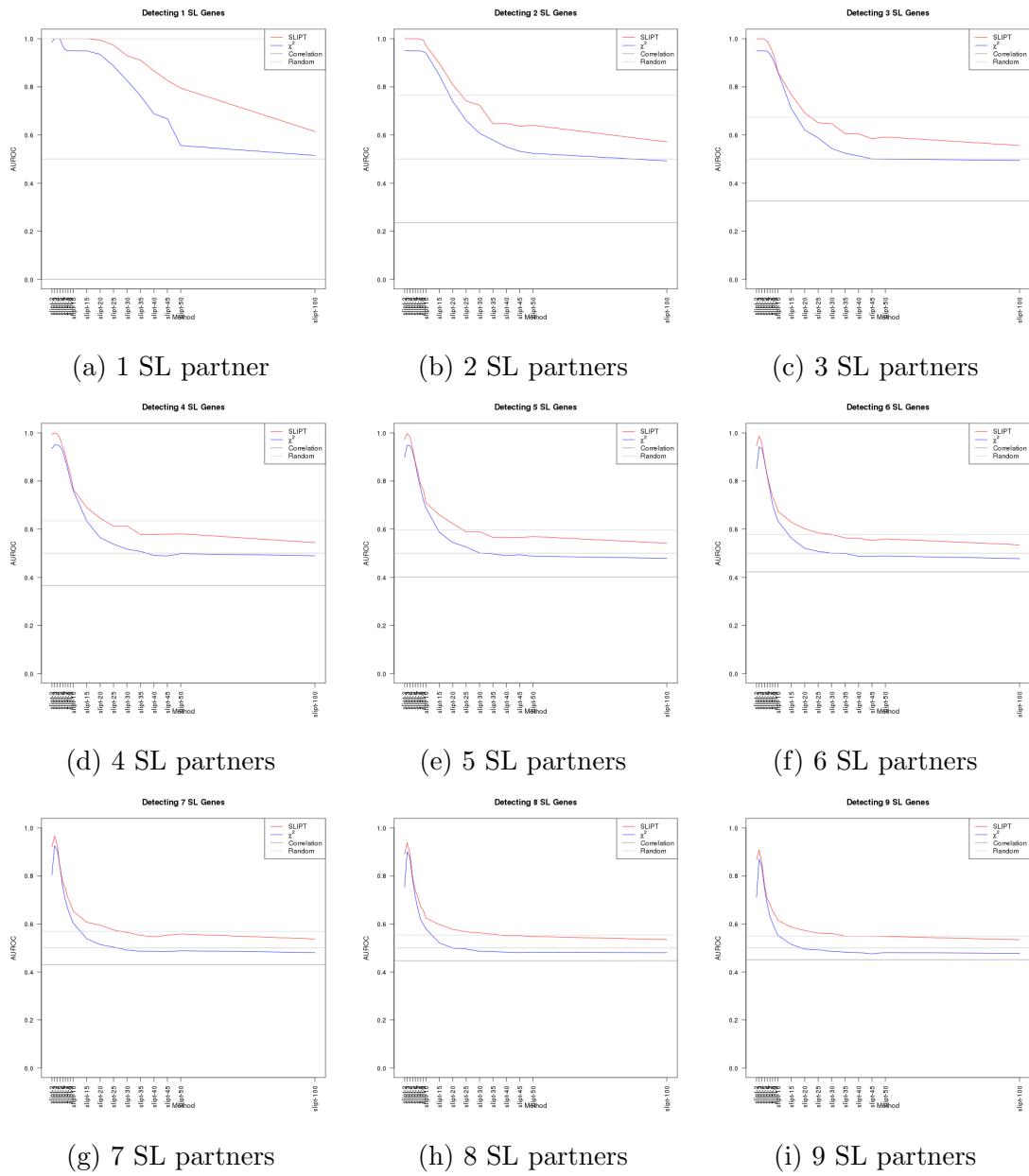


Figure J.6: **Performance of χ^2 and SLIPT across quantiles with query correlation and more genes.** (continued on next page)

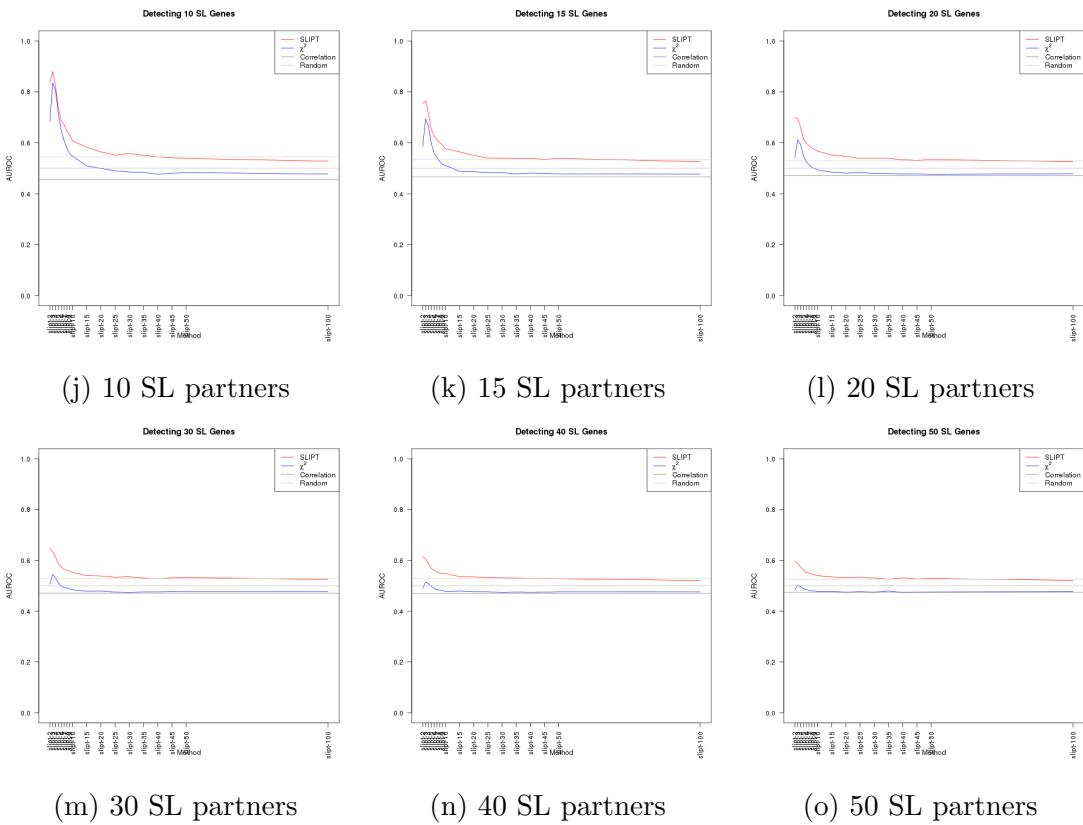
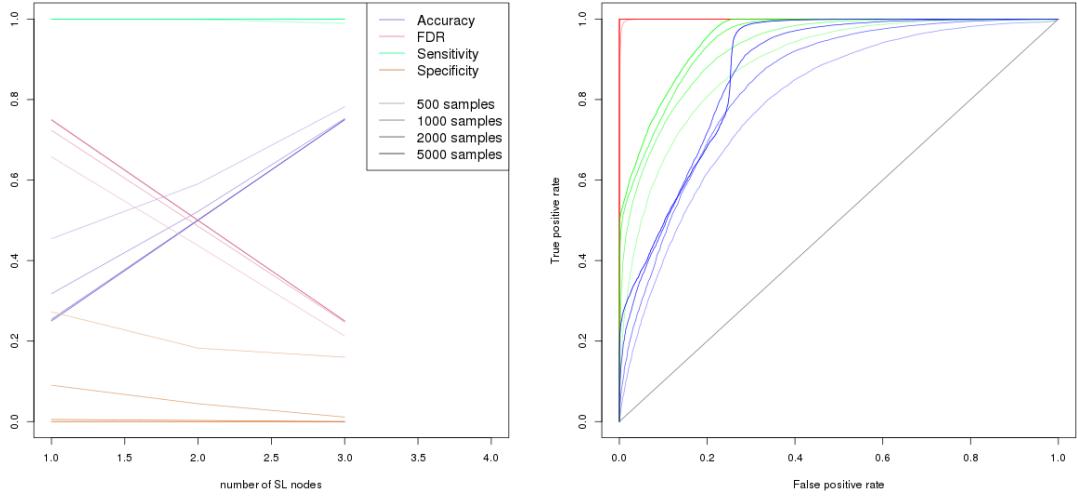


Figure J.6: Performance of χ^2 and SLIPT across quantiles with query correlation and more genes. Synthetic lethal detection with quantiles as in axis labels. The line plots are coloured for SLIPT (red), χ^2 (blue) and correlation (grey) according to the legend. SLIPT and χ^2 perform similarly, peaking at $\frac{1}{3}$ -quantiles and converging to random (0.5). Negative correlation was higher than positive but not optimal quantiles for SLIPT or χ^2 . These findings are robust across different numbers of underlying synthetic lethal genes in 1000 simulations of 20,000 genes (including 1000 correlated with the query) and 1000 samples. SLIPT performs consistently better than χ^2 with positively correlated genes.

Appendix K

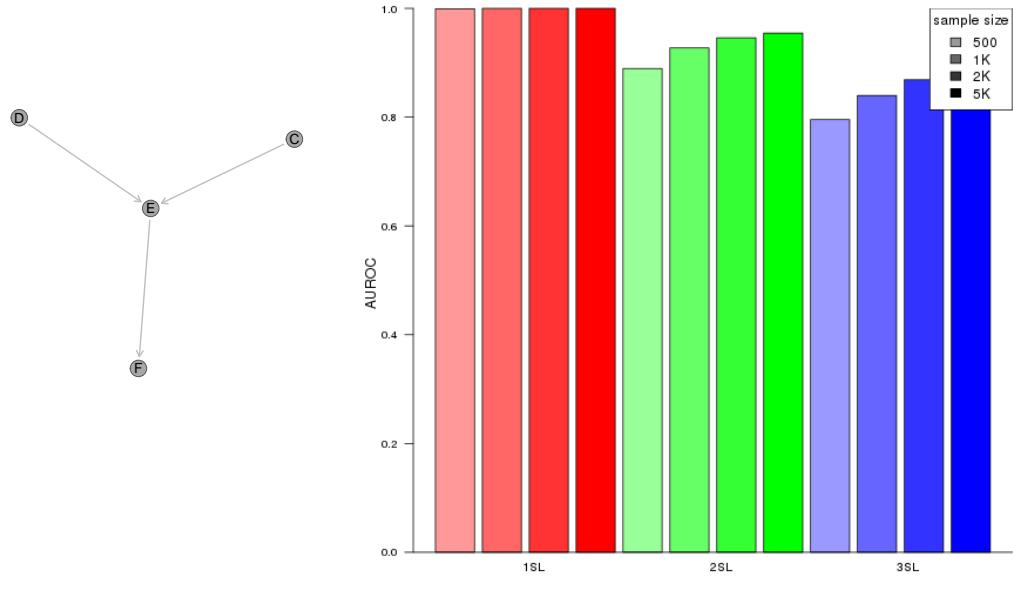
Graph Structures

K.1 Simulations from Simple Graph Structures



(a) Statistical evaluation

(b) Receiver operating characteristic



(c) Graph Structure

(d) Statistical performance

Figure K.1: **Performance of simulations on a simple graph.** Simulation of synthetic lethality was performed using a multivariate normal distribution from a converging graph. For each parameter, 10,000 simulations were used. Colours in Figure K.1b match Figure K.1d.

K.1.1 Simulations from Inhibiting Graph Structures

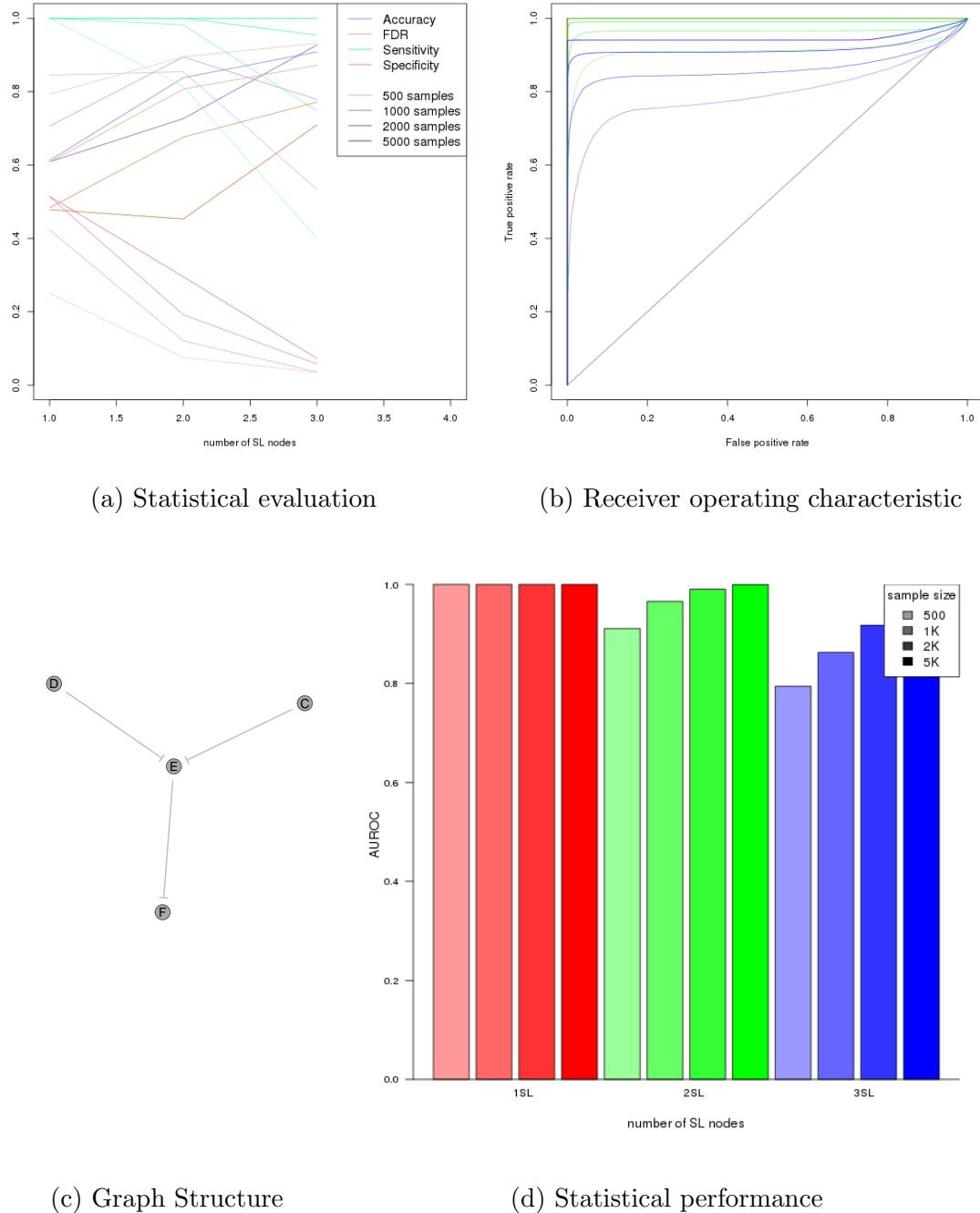
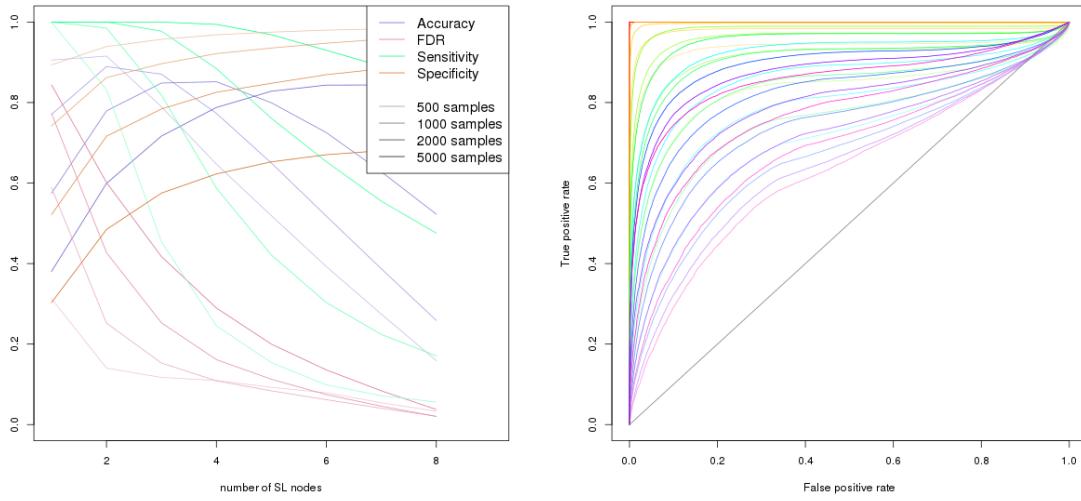
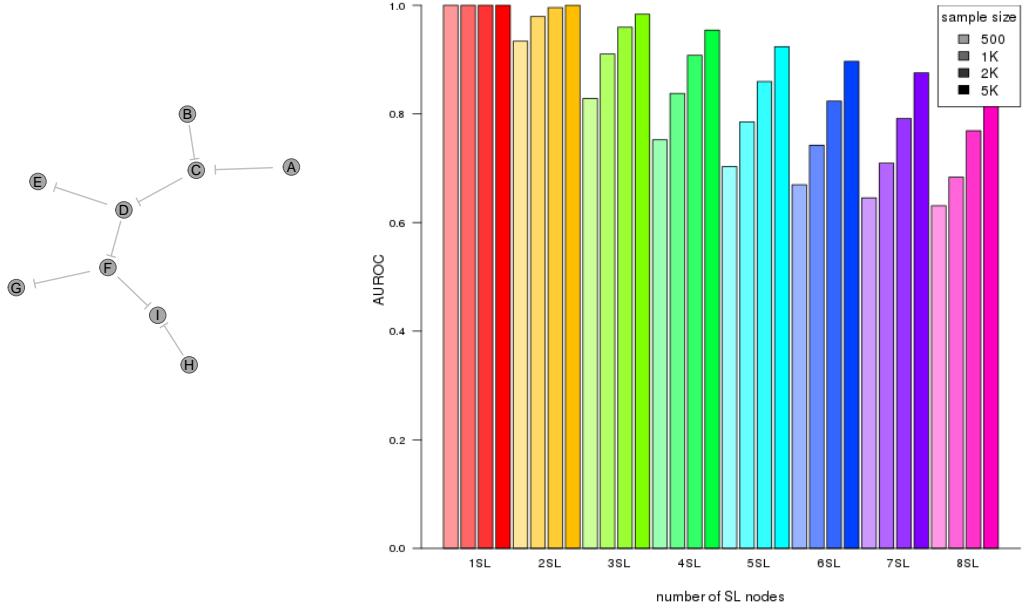


Figure K.2: Performance of simulations on an inhibiting graph. Simulation of synthetic lethality used a multivariate normal distribution from a converging graph. For each parameter, 10,000 simulations were used. Colours in Figure K.2b match Figure K.2d.



(a) Statistical evaluation

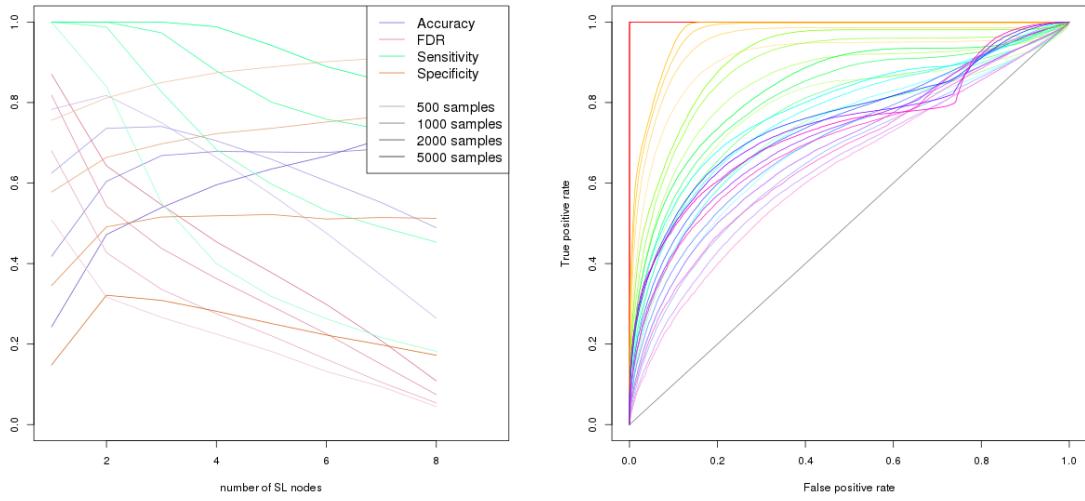
(b) Receiver operating characteristic



(c) Graph Structure

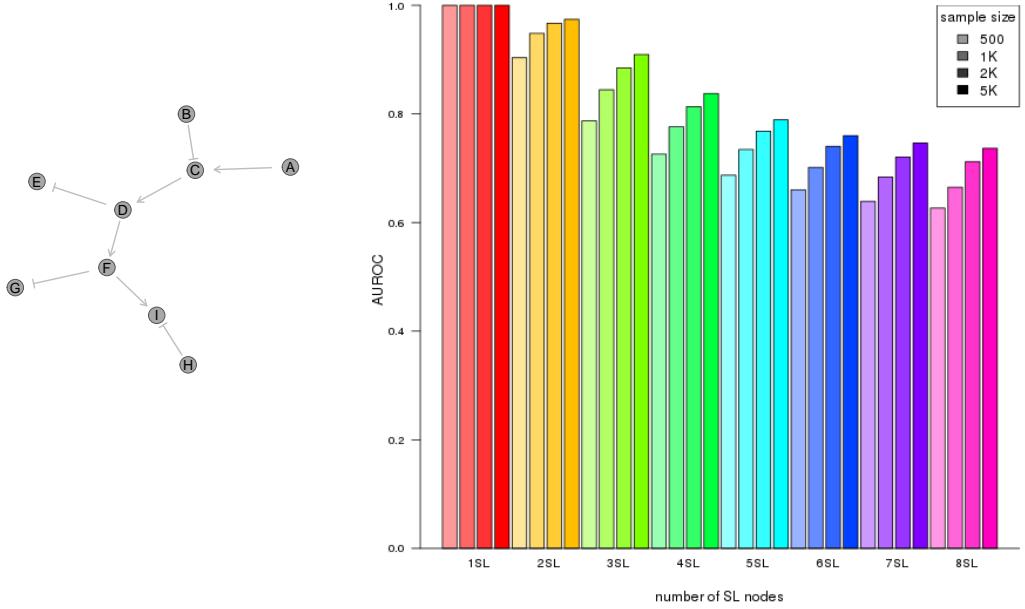
(d) Statistical performance

Figure K.3: Performance of simulations on a constructed graph with inhibition.
 Simulation of synthetic lethality used a multivariate normal distribution from Graph4 with only inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation

(b) Receiver operating characteristic

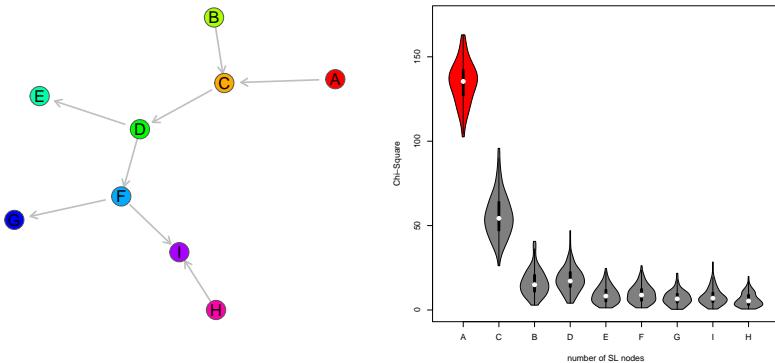


(c) Graph Structure

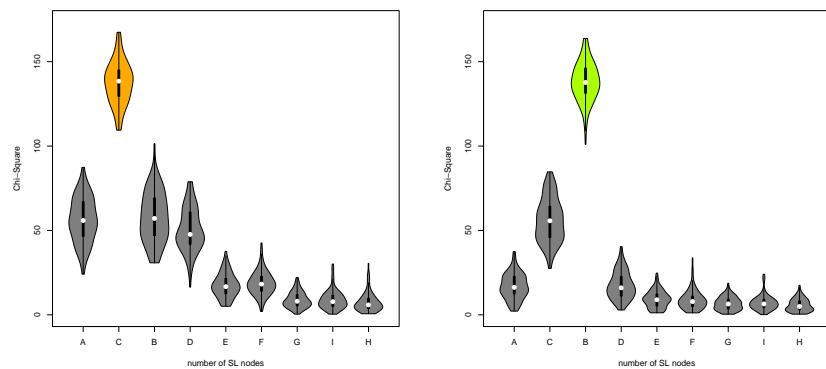
(d) Statistical performance

Figure K.4: Performance of simulations on a constructed graph with inhibition.
 Simulation of synthetic lethality used a multivariate normal distribution from Graph4 with a combination of inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.

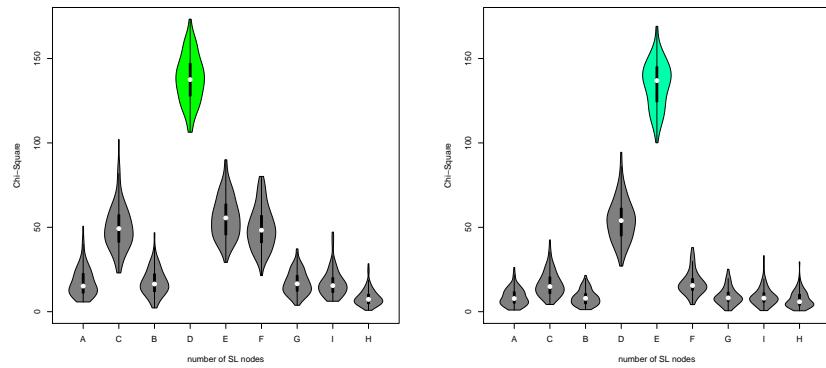
K.2 Simulation across Graph Structures



(a) Activating Graph Structure (b) χ^2 distribution for "A" SL



(c) Gene "B" SL (d) Gene "C" SL



(e) Gene "D" SL (f) Gene "E" SL

Figure K.5: **Detection of Synthetic Lethality within a Graph Structure.** (continued on next page)

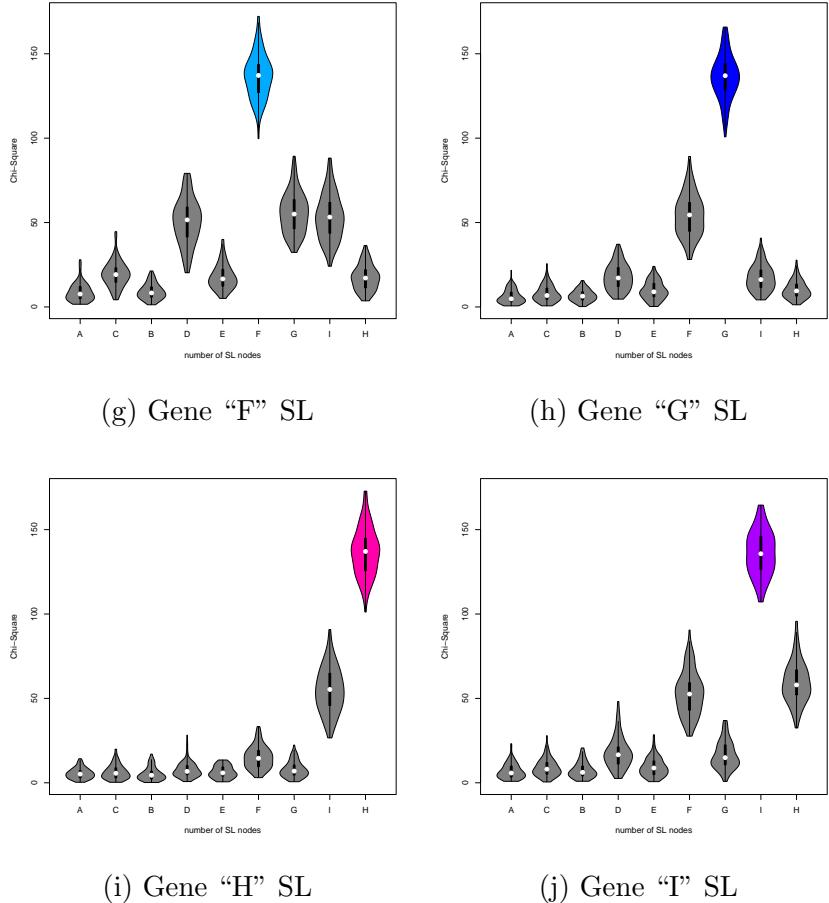
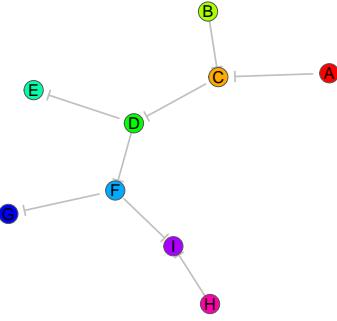
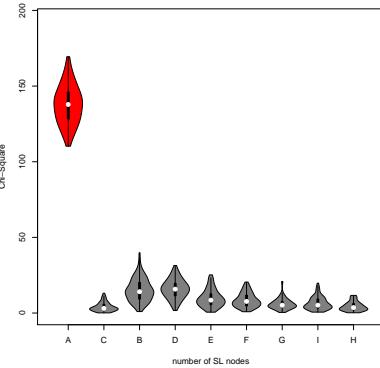


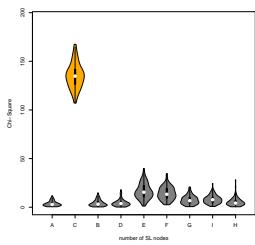
Figure K.5: Detection of Synthetic Lethality within a Graph Structure. Each gene was designated to be synthetic lethal separately and the χ^2 value from SLIPT was computed for each gene across the graph. For each synthetic lethal gene (highlighted in the respective colours), the χ^2 values were computed in 100 simulations of datasets of 20,000 genes including the graph structure and 1000 samples. For each synthetic lethal gene, the adjacent genes in the network also had elevated test statistics.



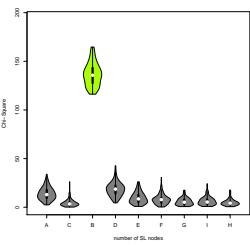
(a) Inhibiting Graph Structure



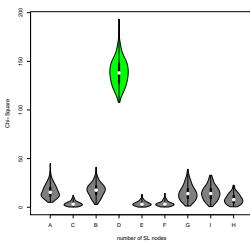
(b) χ^2 distribution for "A" SL



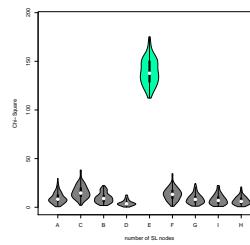
(c) Gene "B" SL



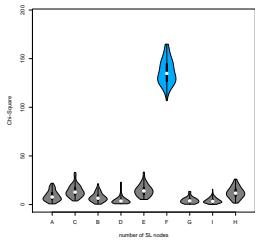
(d) Gene "C" SL



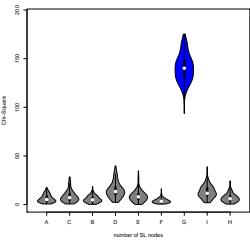
(e) Gene "D" SL



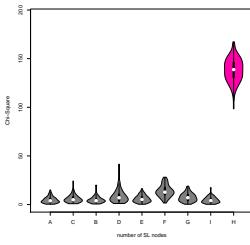
(f) Gene "E" SL



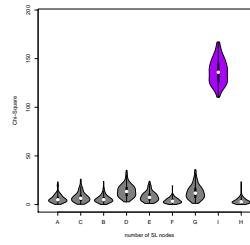
(g) Gene "F" SL



(h) Gene "G" SL

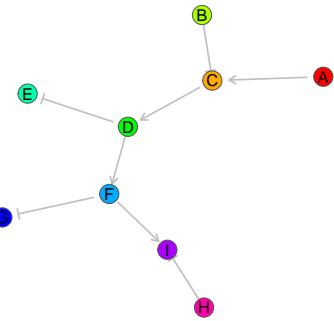


(i) Gene "H" SL

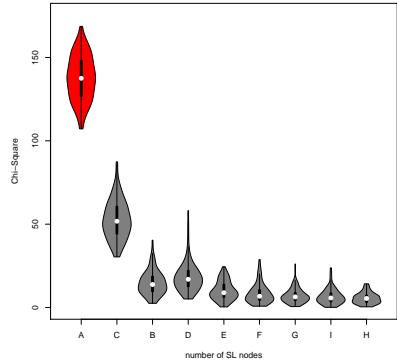


(j) Gene "I" SL

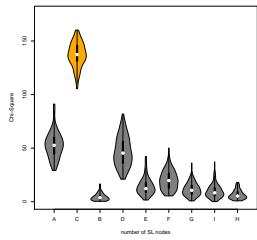
Figure K.6: Detection of Synthetic Lethality within an Inhibiting Graph Structure. Each gene was designated to be synthetic lethal separately and the χ^2 value from SLIPT was computed for each gene across the graph structure with inhibiting relationships. For each synthetic lethal gene (highlighted in the respective colours), the χ^2 values were computed in 100 simulations of datasets of 20,000 genes including the graph structure and 1000 samples. For each synthetic lethal gene, the adjacent genes exhibited lower χ^2 values with inhibiting relationships.



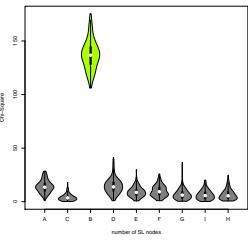
(a) Inhibiting Graph Structure



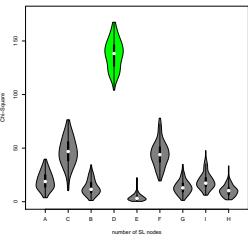
(b) χ^2 distribution for "A" SL



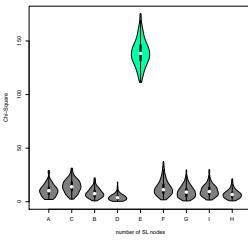
(c) Gene "B" SL



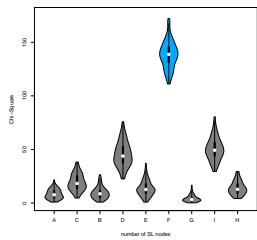
(d) Gene "C" SL



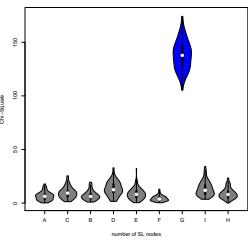
(e) Gene "D" SL



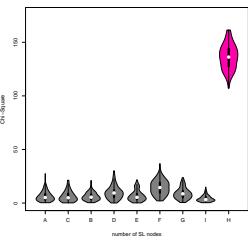
(f) Gene "E" SL



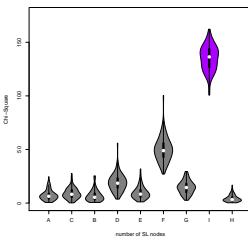
(g) Gene "F" SL



(h) Gene "G" SL



(i) Gene "H" SL



(j) Gene "I" SL

Figure K.7: Detection of Synthetic Lethality within an Inhibiting Graph Structure. Each gene was designated to be synthetic lethal separately and the χ^2 value from SLIPT was computed for each gene across the graph structure with inhibiting and relationships. For each synthetic lethal gene (highlighted in the respective colours), the χ^2 values were computed in 100 simulations of datasets of 20,000 genes including the graph structure and 1000 samples.

K.3 Simulations from Complex Graph Structures

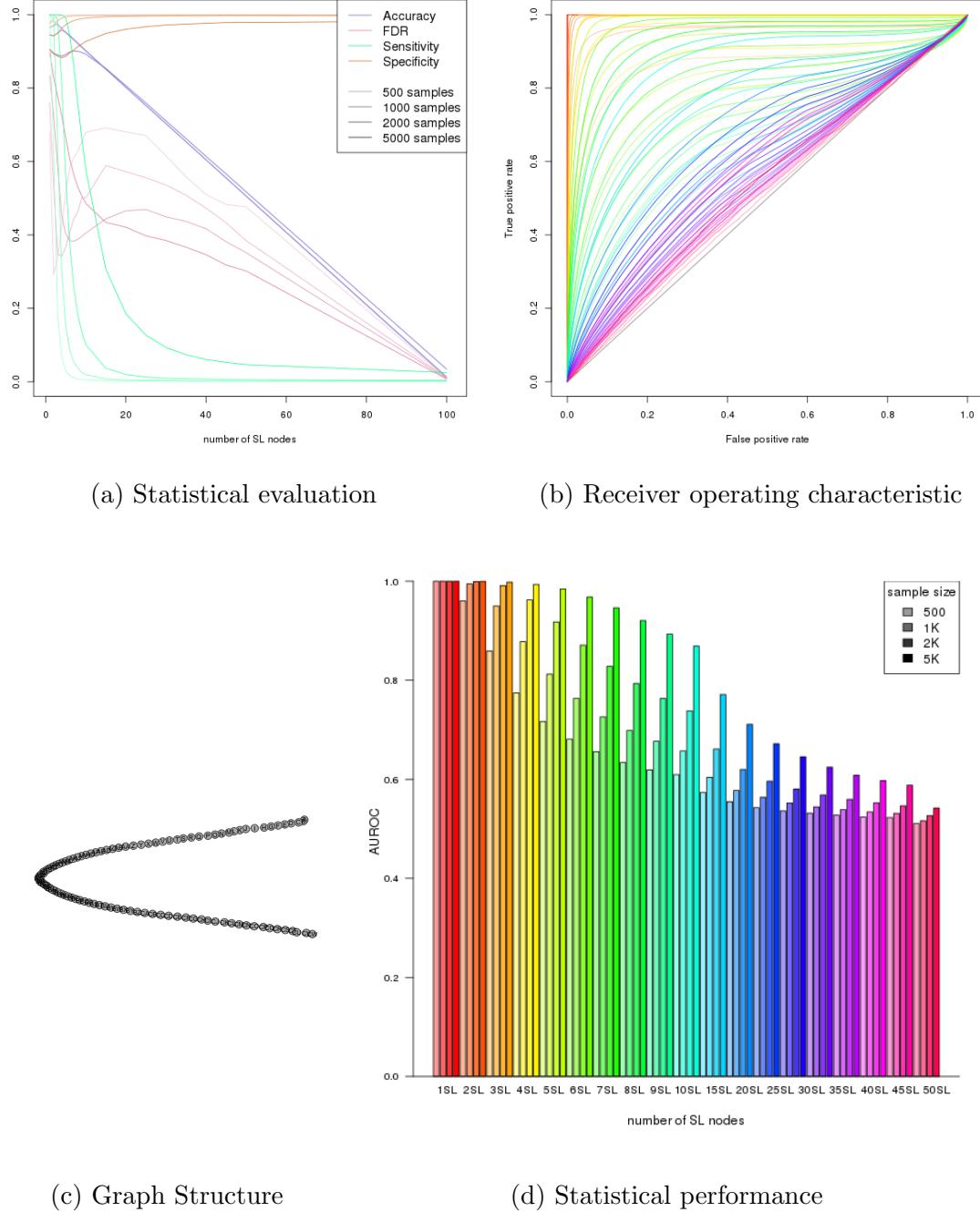
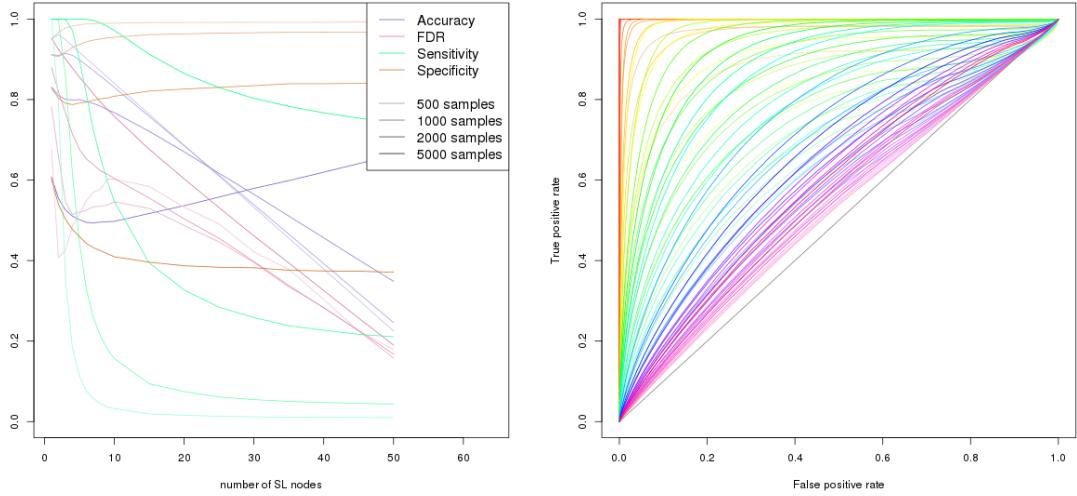
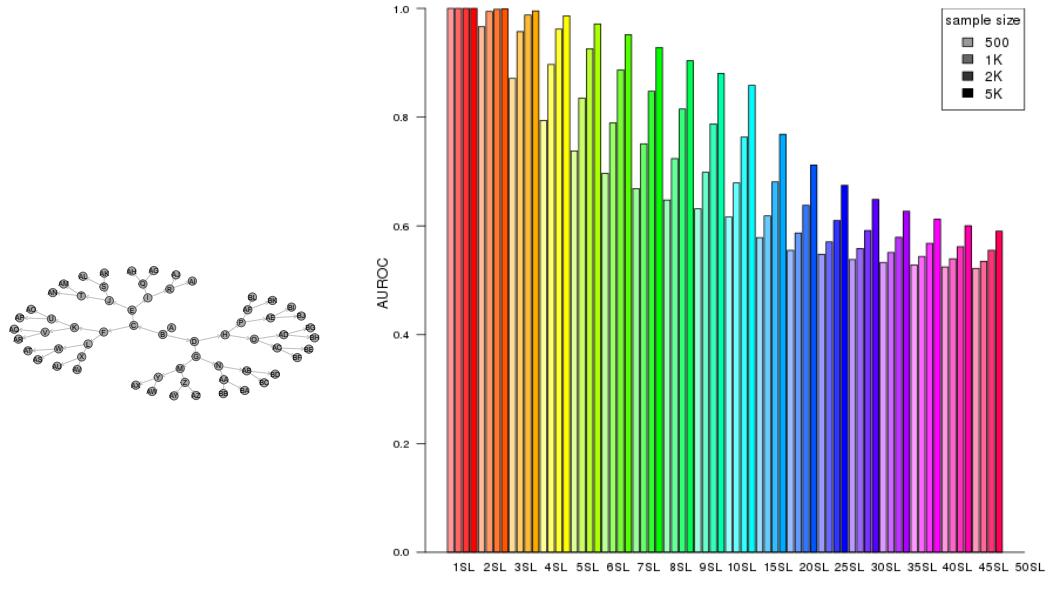


Figure K.8: **Performance of simulations on a large graph.** Simulation of synthetic lethality used a multivariate normal distribution from a large graph. For each parameter, 10,000 simulations were used. Colours in Figure K.8b match Figure K.8d.



(a) Statistical evaluation

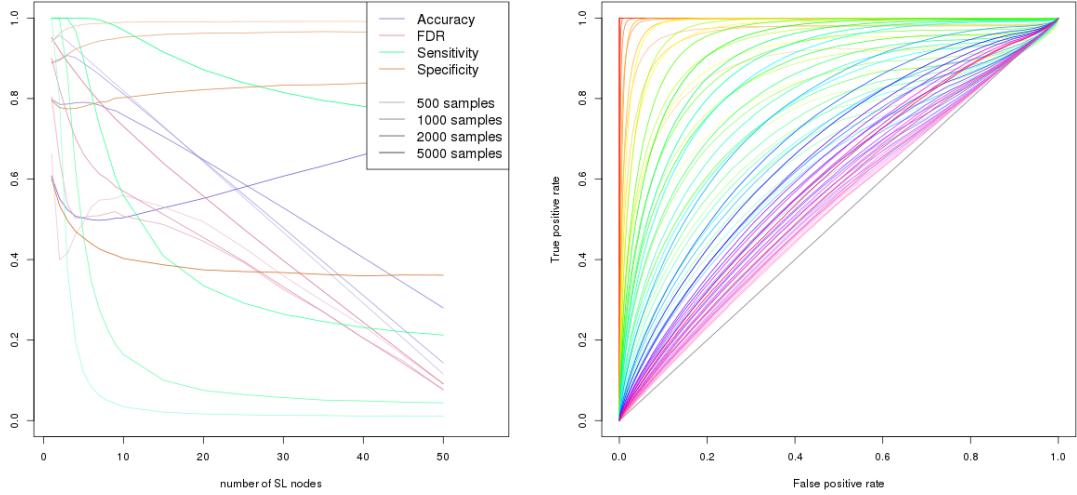
(b) Receiver operating characteristic



(c) Graph Structure

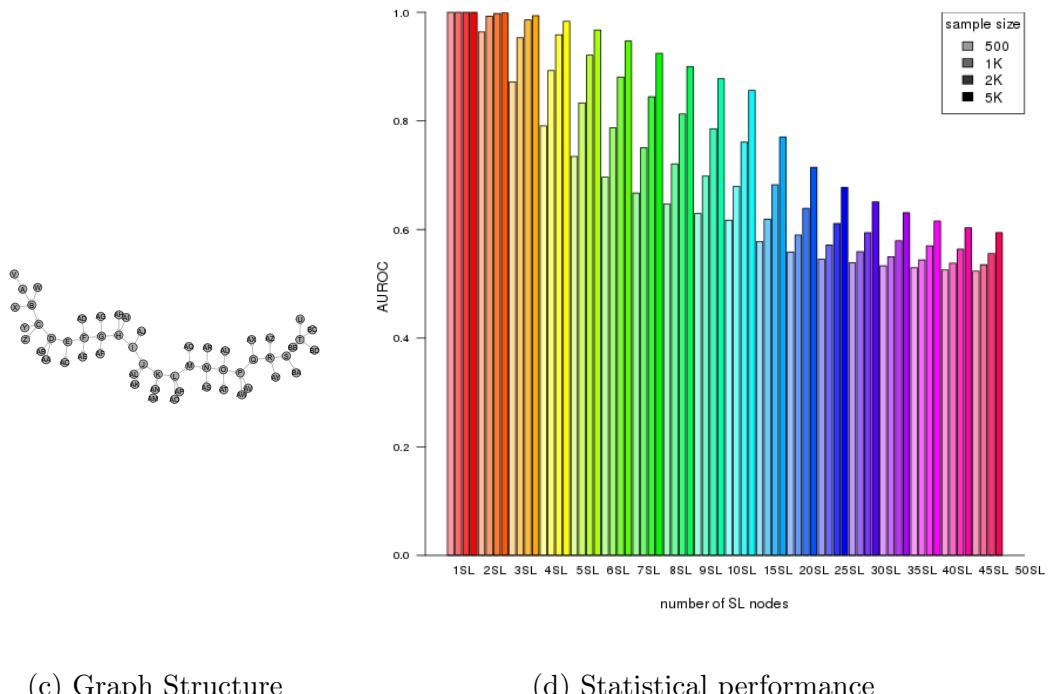
(d) Statistical performance

Figure K.9: Performance of simulations on a branching graph. Simulation of synthetic lethality used a multivariate normal distribution from a branching graph. For each parameter, 10,000 simulations were used. Colours in Figure K.9b match Figure K.9d.



(a) Statistical evaluation

(b) Receiver operating characteristic

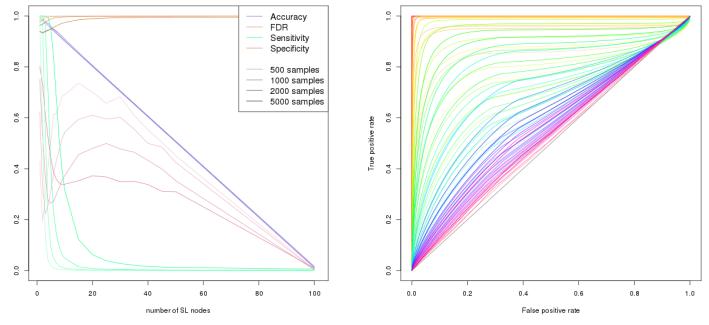


(c) Graph Structure

(d) Statistical performance

Figure K.10: Performance of simulations on a complex graph. Simulation of synthetic lethality used a multivariate normal distribution from a complex graph. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used. Colours in Figure K.10b match Figure K.10d.

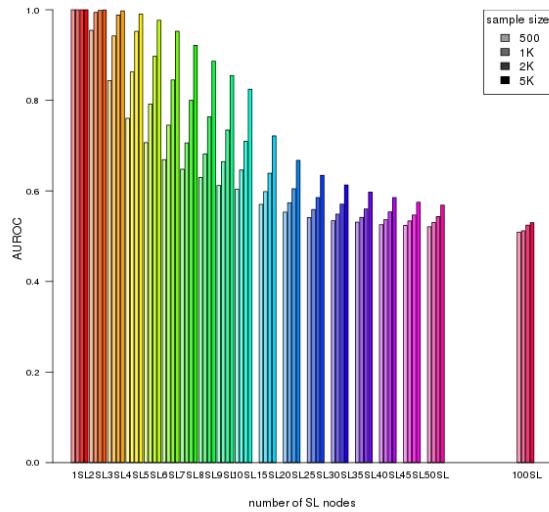
K.3.1 Simulations from Complex Inhibiting Graphs



(a) Statistical evaluation (b) Receiver operating characteristic

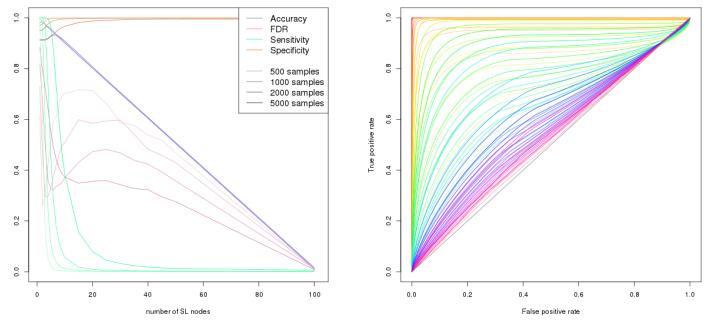


(c) Graph Structure

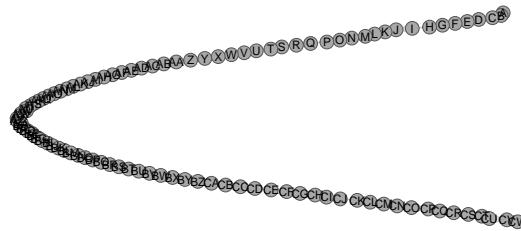


(d) Statistical performance

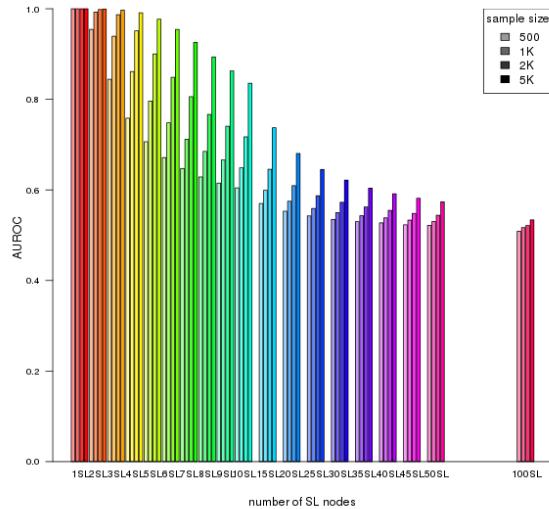
Figure K.11: Performance of simulations on a large constructed graph with inhibition. Simulation of synthetic lethality used a multivariate normal distribution from Graph5 with only inhibitions. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation (b) Receiver operating characteristic

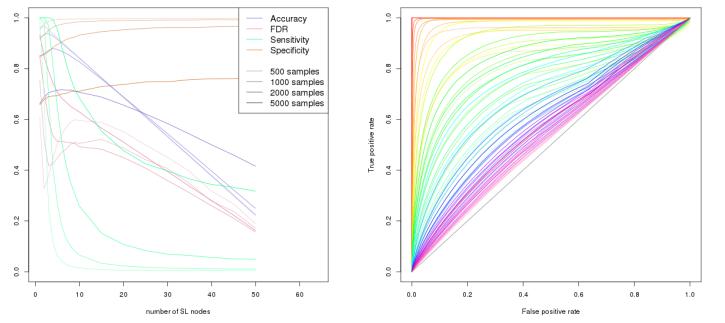


(c) Graph Structure

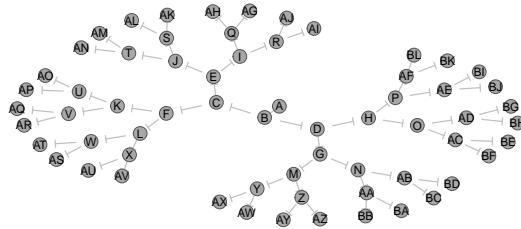


(d) Statistical performance

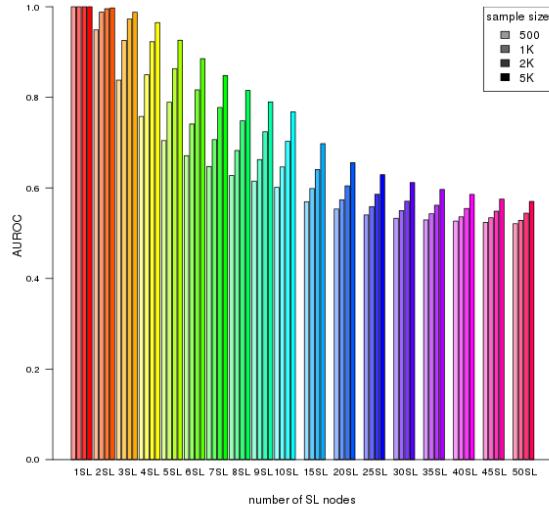
Figure K.12: Performance of simulations on a large constructed graph with inhibition. Simulation of synthetic lethality used a multivariate normal distribution from Graph5 with alternating inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation (b) Receiver operating characteristic

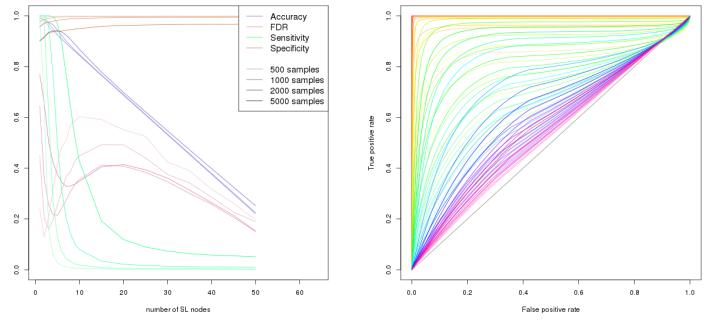


(c) Graph Structure

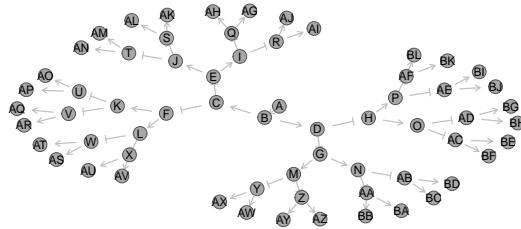


(d) Statistical performance

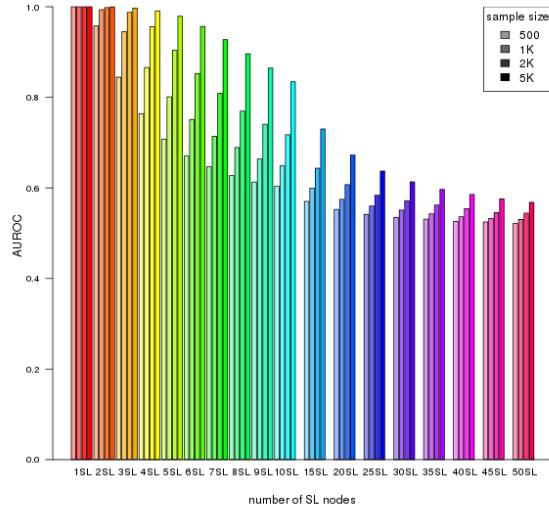
Figure K.13: Performance of simulations on a branching graph with inhibition.
 Simulation of synthetic lethality used a multivariate normal distribution from Graph6 with only inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation (b) Receiver operating characteristic

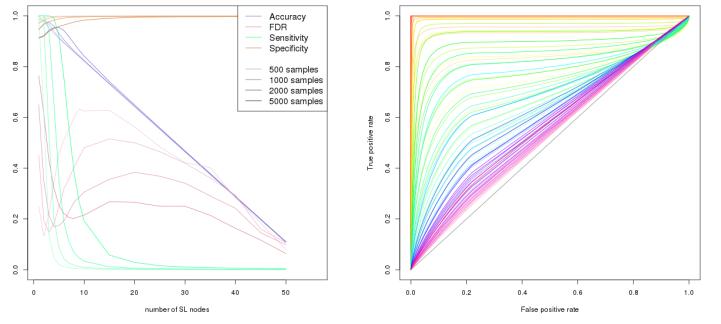


(c) Graph Structure

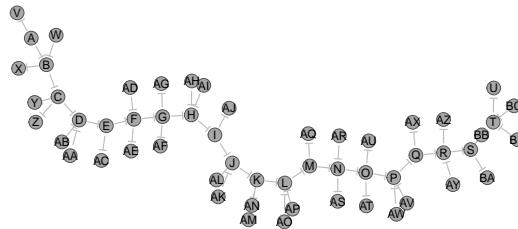


(d) Statistical performance

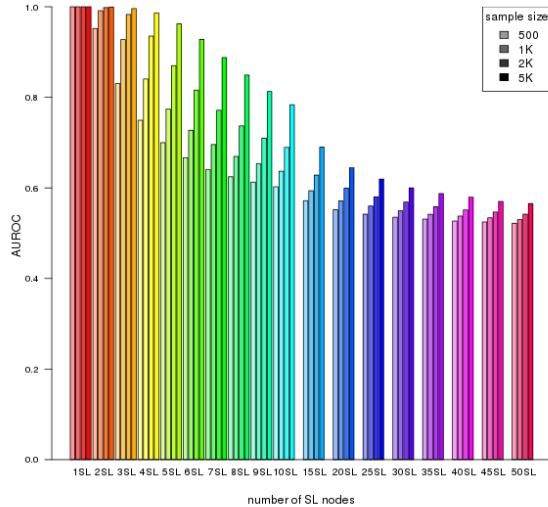
Figure K.14: Performance of simulations on a branching graph with inhibition.
 Simulation of synthetic lethality used a multivariate normal distribution from Graph6 with alternating inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation (b) Receiver operating characteristic

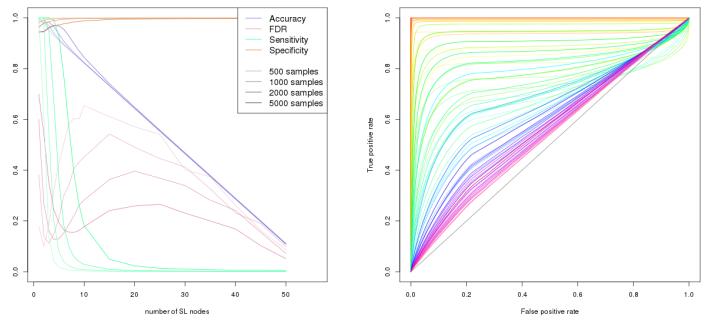


(c) Graph Structure

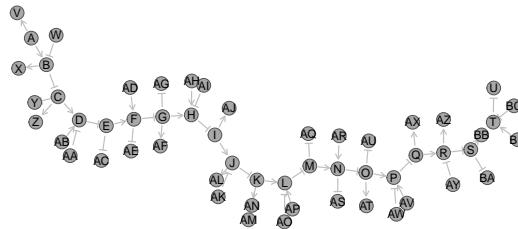


(d) Statistical performance

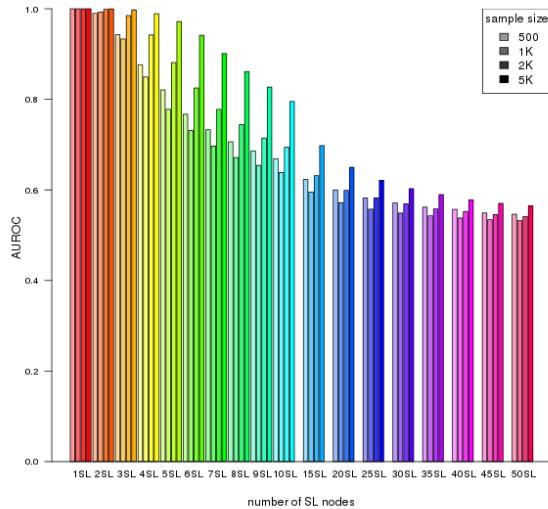
Figure K.15: **Performance of simulations on a complex graph with inhibition.** Simulation of synthetic lethality used a multivariate normal distribution from Graph7 with only inhibitions. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.



(a) Statistical evaluation (b) Receiver operating characteristic



(c) Graph Structure



(d) Statistical performance

Figure K.16: Performance of simulations on a complex graph with inhibition. Simulation of synthetic lethality used a multivariate normal distribution from Graph7 with a combination of relationships. Performance of SLIPT declines for more synthetic partners and lower sample sizes. For each parameter, 10,000 simulations were used.

K.4 Simulations from Pathway Graph Structures

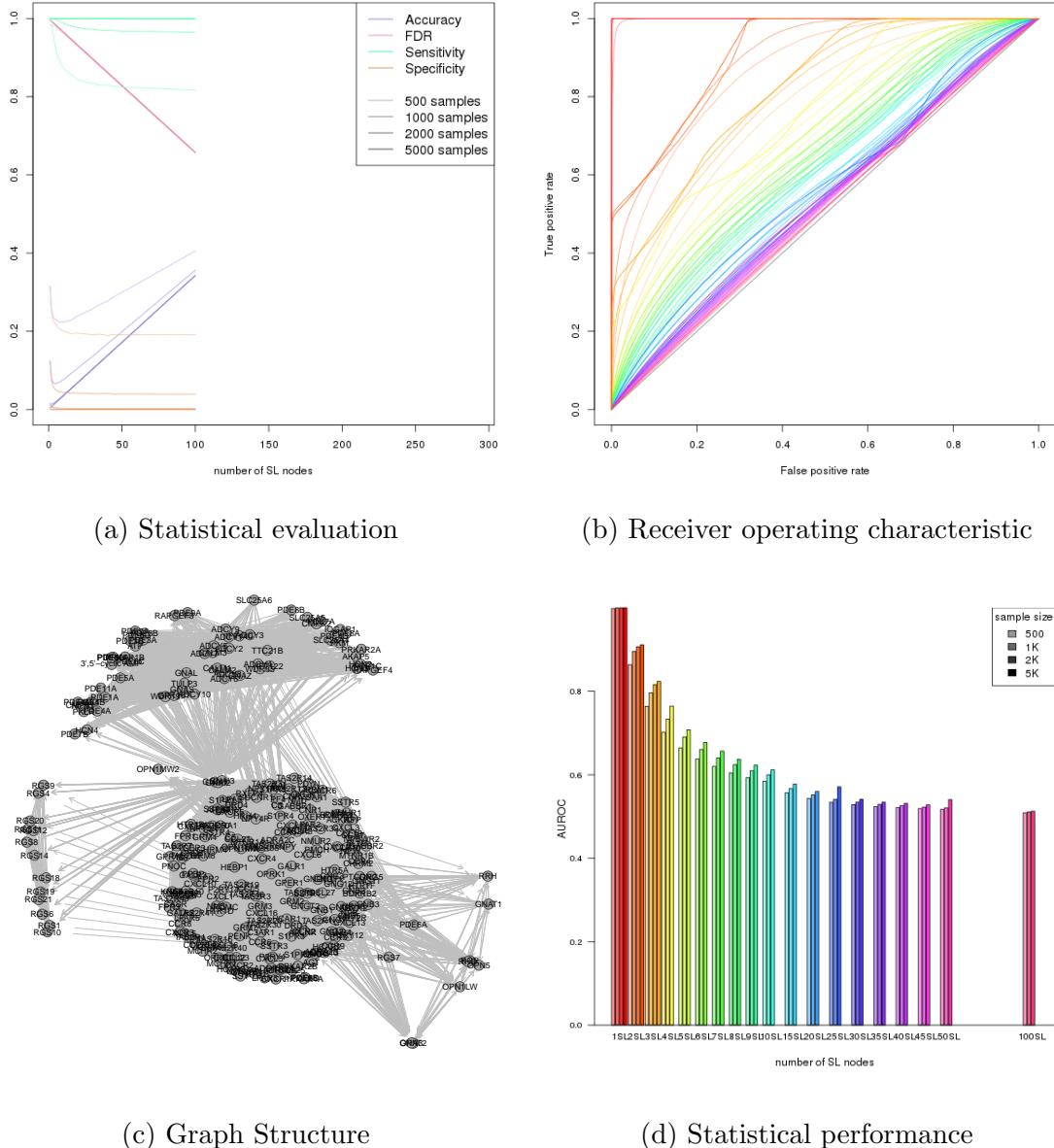


Figure K.17: Performance of simulations on the $G_{\alpha i}$ signalling pathway. Simulation of synthetic lethality used a multivariate normal distribution based on the Reactome $G_{\alpha i}$ signalling pathway. Performance of SLIPT was high across parameters for detecting synthetic lethality in the 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.

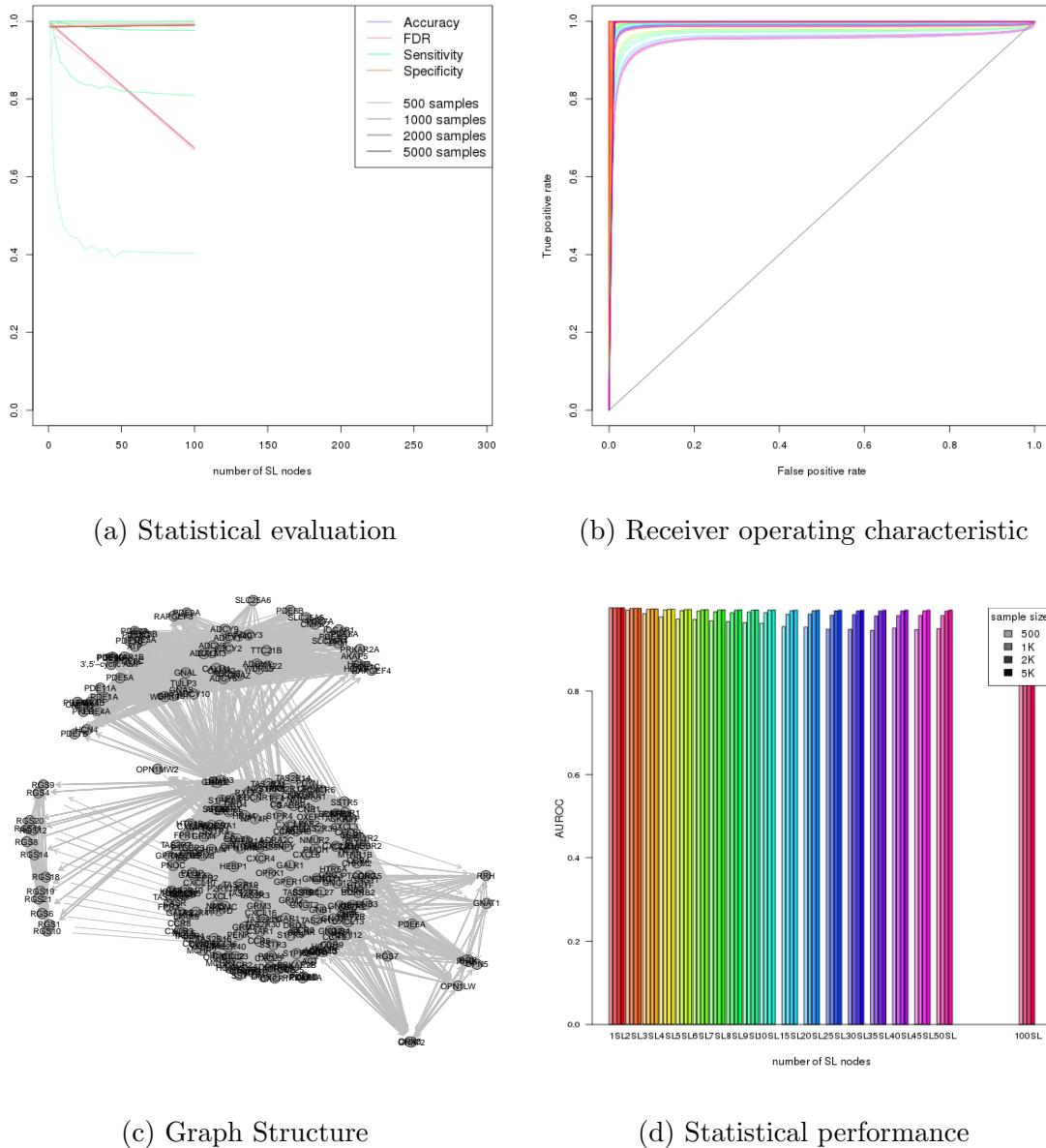


Figure K.18: Performance of simulations including the $G_{\alpha i}$ signalling pathway. Simulation of synthetic lethality used a multivariate normal distribution (without correlation structure apart from the Reactome $G_{\alpha i}$ signalling pathway. Performance of SLIPT was high across parameters 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.