

Contents

Glossary	xiv
Acronyms	xv
1 Introduction	1
1.1 Cancer Research in the Post-Genomic Era	1
1.1.1 Cancer as a Global Health Concern	2
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	11
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	13
1.1.5 Cancer Genomes	14
1.1.5.1 The Cancer Genome Atlas Project	15
1.1.5.1.1 Findings from Cancer Genomes	15
1.1.5.1.2 Genomic Comparisons Across Cancer Tissues	17
1.1.5.1.3 Cancer Genomic Data Resources	18
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	21
1.1.6.3.1 Targeting Oncogenic Driver Mutations	21
1.1.6.4 Systems and Network Biology	22
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	26
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	28
1.2.3.1.1	Evolution of Synthetic Lethality	29
1.2.4	Synthetic Lethal Concepts in Cancer	29
1.2.5	Clinical Impact of Synthetic Lethality in Cancer	31
1.2.6	High-throughput Screening for Synthetic Lethality	33
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	43
1.2.7.5	Data Mining and Machine Learning	44
1.2.7.6	Bimodality	47
1.2.7.7	Rationale for Further Development	48
1.3	E-cadherin as a Synthetic Lethal Target	48
1.3.1	The <i>CDH1</i> gene and it's Biological Functions	48
1.3.1.1	Cytoskeleton	49
1.3.1.2	Extracellular and Tumour Micro-Environment	49
1.3.1.3	Cell-Cell Adhesion and Signalling	49
1.3.2	<i>CDH1</i> as a Tumour (and Invasion) Suppressor	50
1.3.2.1	Breast Cancers and Invasion	50
1.3.3	Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer .	50
1.3.4	Somatic Mutations	52
1.3.4.1	Mutation Rate	52
1.3.4.2	Co-occurring Mutations	52
1.3.5	Models of <i>CDH1</i> loss in cell lines	53
1.4	Summary and Research Direction of Thesis	54
2	Methods and Resources	58
2.1	Bioinformatics Resources for Genomics Research	58
2.1.1	Public Data and Software Packages	58
2.1.1.1	Cancer Genome Atlas Data	59
2.1.1.2	Reactome and Annotation Data	60
2.2	Data Handling	61
2.2.1	Normalisation	61
2.2.2	Sample Triage	61
2.2.3	Metagenes and the Singular Value Decomposition	63
2.2.3.1	Candidate Triage and Integration with Screen Data . .	63
2.3	Techniques	64
2.3.1	Statistical Procedures and Tests	64
2.3.2	Gene Set Over-representation Analysis	65
2.3.3	Clustering	66
2.3.4	Heatmap	66
2.3.5	Modeling and Simulations	66
2.3.5.1	Receiver Operating Characteristic (Performance) . . .	67
2.3.6	Resampling Analysis	68

2.4	Pathway Structure Methods	69
2.4.1	Network and Graph Analysis	69
2.4.2	Sourcing Graph Structure Data	70
2.4.3	Constructing Pathway Subgraphs	70
2.4.4	Network Analysis Metrics	70
2.5	Implementation	71
2.5.1	Computational Resources and Linux Utilities	71
2.5.2	R Language and Packages	73
2.5.3	High Performance and Parallel Computing	75
3	Methods Developed During Thesis	77
3.1	A Synthetic Lethal Detection Methodology	77
3.2	Synthetic Lethal Simulation and Modelling	80
3.2.1	A Model of Synthetic Lethality in Expression Data	80
3.2.2	Simulation Procedure	84
3.3	Detecting Simulated Synthetic Lethal Partners	87
3.3.1	Binomial Simulation of Synthetic lethality	87
3.3.2	Multivariate Normal Simulation of Synthetic lethality	89
	3.3.2.1 Multivariate Normal Simulation with Correlated Genes	92
	3.3.2.2 Specificity with Query-Correlated Pathways	99
	3.3.2.2.1 Importance of Directional Testing	99
3.4	Graph Structure Methods	101
3.4.1	Upstream and Downstream Gene Detection	101
	3.4.1.1 Permutation Analysis for Statistical Significance	102
	3.4.1.2 Hierarchy Based on Biological Context	103
3.4.2	Simulating Gene Expression from Graph Structures	104
3.5	Customised Functions and Packages Developed	108
3.5.1	Synthetic Lethal Interaction Prediction Tool	108
3.5.2	Data Visualisation	109
3.5.3	Extensions to the iGraph Package	111
	3.5.3.1 Sampling Simulated Data from Graph Structures	111
	3.5.3.2 Plotting Directed Graph Structures	111
	3.5.3.3 Computing Information Centrality	112
	3.5.3.4 Testing Pathway Structure with Permutation Testing	112
	3.5.3.5 Metapackage to Install iGraph Functions	113
4	Synthetic Lethal Analysis of Gene Expression Data	114
4.1	Synthetic lethal genes in breast cancer	115
4.1.1	Synthetic lethal pathways in breast cancer	117
4.1.2	Expression profiles of synthetic lethal partners	118
	4.1.2.1 Subgroup pathway analysis	121
4.2	Comparison of synthetic lethal gene candidates	124
4.2.1	Comparison with siRNA screen candidates	124
	4.2.1.1 Comparison with correlation	125
	4.2.1.2 Comparison with viability	126
	4.2.1.3 Comparison with secondary siRNA screen candidates	130

4.2.1.4	Comparison of screen at pathway level	130
4.2.1.4.1	Resampling of genes for pathway enrichment . .	132
4.3	Metagene Analysis	138
4.3.1	Pathway expression	138
4.3.2	Somatic mutation	141
4.3.3	Mutation locus	142
4.3.4	Synthetic lethal metagenes	144
4.4	Replication in stomach cancer	145
4.4.1	Synthetic Lethal Genes and Pathways	146
4.4.2	Synthetic Lethal Expression Profiles	148
4.4.3	Comparison to Primary Screen	150
4.4.3.1	Resampling Analysis	151
4.4.4	Metagene Analysis	151
4.5	Global Synthetic Lethality	152
4.5.1	Hub Genes	153
4.5.2	Hub Pathways	155
4.6	Replication in cell line encyclopaedia	156
4.7	Discussion	158
4.7.1	Strengths of the SLIPT Methodology	158
4.7.2	Synthetic Lethal Pathways for E-cadherin	159
4.7.3	Replication and Validation	161
4.7.3.1	Integration with siRNA Screening	161
4.7.3.2	Replication across Tissues and Cell lines	162
4.8	Summary	163
5	Synthetic Lethal Pathway Structure	166
5.1	Synthetic Lethal Genes in Reactome Pathways	167
5.1.1	The PI3K/AKT Pathway	167
5.1.2	The Extracellular Matrix	169
5.1.3	G Protein Coupled Receptors	172
5.1.4	Gene Regulation and Translation	172
5.2	Network Analysis of Synthetic Lethal Genes	173
5.2.1	Gene Connectivity and Vertex Degree	173
5.2.2	Gene Importance and Centrality	175
5.2.2.1	Information Centrality	175
5.2.2.2	PageRank Centrality	178
5.3	Testing Pathway Structure of Synthetic Lethal Genes	179
5.3.1	Hierarchical Pathway Structure	179
5.3.1.1	Contextual Hierarchy of PI3K	179
5.3.1.2	Testing Contextual Hierarchy of Synthetic Lethal Genes	179
5.3.2	Upstream or Downstream Synthetic Lethality	183
5.3.2.1	Measuring Structure of Candidates within PI3K	183
5.3.2.2	Resampling for Synthetic Lethal Pathway Structure . .	185
5.4	Discussion	186
5.5	Summary	188

6	Simulation and Modeling of Synthetic Lethal Pathways	191
6.1	Comparing methods	192
6.1.1	Performance of SLIPT and χ^2 across Quantiles	193
6.1.1.1	Correlated Query Genes affects Specificity	196
6.1.2	Alternative Synthetic Lethal Detection Strategies	198
6.1.2.1	Correlation for Synthetic Lethal Detection	198
6.1.2.2	Testing for Bimodality with BiSEp	200
6.2	Simulations with Graph Structures	202
6.2.1	Performance over a Graph Structure	202
6.2.1.1	Simple Graph Structures	202
6.2.1.2	Constructed Graph Structures	202
6.2.2	Performance with Inhibitions	207
6.2.3	Synthetic Lethality across Graph Structures	212
6.2.4	Performance with Feasible Gene Numbers (20,000)	216
6.2.4.1	Simple Graph Structures in a Genome	216
6.2.4.2	Constructed Graph Structures in a Genome	216
6.3	Simulations over pathway-based graphs	225
6.3.1	Pathway Structures in a Genome	227
6.4	Discussion	228
6.5	Summary	228
7	Discussion	228
7.1	Significance	228
7.2	Future Directions	229
7.3	Conclusion	230
8	Conclusion	234
	References	235
A	Sample Quality	260
A.1	Sample Correlation	260
A.2	Replicate Samples in TCGA Breast	263
B	Software Used for Thesis	267
C	Secondary Screen Data	276
D	Mutation Analysis in Breast Cancer	278
D.1	Synthetic Lethal Genes and Pathways	278
D.2	Synthetic Lethal Expression Profiles	281
D.3	Comparison to Primary Screen	284
D.3.1	Resampling Analysis	286
D.4	Compare SLIPT genes	288
D.5	Metagene Analysis	290
D.6	Mutation Variation	291
D.6.1	Mutation Frequency	291

D.6.2	PI3K Mutation Expression	292
E	Metagene Expression Profiles	295
F	Stomach Expression Analysis	301
F.1	Synthetic Lethal Genes and Pathways	301
F.2	Comparison to Primary Screen	304
F.2.1	Resampling Analysis	306
F.3	Metagene Analysis	308
G	Stomach Mutation Analysis	309
G.1	Synthetic Lethal Genes and Pathways	309
G.2	Synthetic Lethal Expression Profiles	312
G.3	Comparison to Primary Screen	315
G.3.1	Resampling Analysis	317
G.4	Metagene Analysis	319
H	Global Synthetic Lethality in Stomach Cancer	320
H.1	Hub Genes	322
H.2	Hub Pathways	323
I	Replication in cell line encyclopaedia	324
J	Synthetic Lethal Genes in Pathways	329
K	Pathway Connectivity for Mutation SLIPT	337
L	Information Centrality for Gene Essentiality	341
M	Pathway Structure for Mutation SLIPT	344
N	Performance of SLIPT and χ^2	347
N.0.1	Correlated Query Genes affects Specificity	353
O	Graph Structures	359
O.1	Simulations from Graph Structures	365
O.2	Simulations from Graph Structures with Inhibition	369
O.3	Simulation across Graph Structures	378
O.4	Simulations from Graph Structures with 20K genes	382
O.5	Simulations from Graph Structures with 20K genes with Inhibition	387
O.6	Simations from Pathway Graph Structures	397

List of Figures

1.1	Synthetic genetic interactions	27
1.2	Synthetic lethality in cancer	30
2.1	Read count density	62
2.2	Read count sample mean	62
3.1	Framework for synthetic lethal prediction	78
3.2	Synthetic lethal prediction adapted for mutation	79
3.3	A model of synthetic lethal gene expression	81
3.4	Modeling synthetic lethal gene expression	82
3.5	Synthetic lethality with multiple genes	83
3.6	Simulating gene function	85
3.7	Simulating synthetic lethal gene function	85
3.8	Simulating synthetic lethal gene expression	86
3.9	Performance of binomial simulations	88
3.10	Comparison of statistical performance	88
3.11	Performance of multivariate normal simulations	90
3.12	Simulating expression with correlated gene blocks	93
3.13	Simulating expression with correlated gene blocks	94
3.14	Synthetic lethal prediction across simulations	95
3.15	Performance with correlations	96
3.16	Comparison of statistical performance with correlation structure	97
3.17	Performance with query correlations	98
3.18	Statistical evaluation of directional criteria	99
3.19	Performance of directional criteria	100
3.20	Simulated graph structures	104
3.21	Simulating expression from a graph structure	106
3.22	Simulating expression from graph structure with inhibitions	107
3.23	Demonstration of violin plots with custom features	110
3.24	Demonstration of annotated heatmap	110
3.25	Simulating graph structures	112
4.1	Synthetic lethal expression profiles of analysed samples	120
4.2	Comparison of SLIPT to siRNA	124
4.3	Compare SLIPT and siRNA genes with correlation	125
4.4	Compare SLIPT and siRNA genes with correlation	125
4.5	Compare SLIPT and siRNA genes with siRNA viability	127

4.6	Compare SLIPT and siRNA genes with viability	127
4.7	Compare SLIPT and siRNA genes with siRNA viability	129
4.8	Resampled intersection of SLIPT and siRNA candidates	133
4.9	Pathway metagene expression profiles	139
4.10	Somatic mutation against PI3K metagene	141
4.11	Somatic mutation locus against expression	143
4.12	Synthetic lethal expression profiles of stomach samples	149
4.13	Synthetic lethal partners across query genes	153
5.1	Synthetic Lethality in the PI3K Cascade	168
5.2	Synthetic Lethality in the Elastic Fibre Formation Pathway	170
5.3	Synthetic Lethality in the Fibrin Clot Formation	171
5.4	Synthetic Lethality and Vertex Degree	174
5.5	Synthetic Lethality and Centrality	176
5.6	Synthetic Lethality and PageRank	178
5.7	Structure of PI3K Ranking	180
5.8	Synthetic Lethality and Hierarchy Score in PI3K	181
5.9	Hierarchy Score in PI3K against Synthetic Lethality in PI3K	181
5.10	Structure of Synthetic Lethality in PI3K	182
5.11	Structure of Synthetic Lethality Resampling in PI3K	184
6.1	Performance of χ^2 and SLIPT across quantiles	194
6.2	Performance of χ^2 and SLIPT across quantiles with more genes	195
6.3	Performance of χ^2 and SLIPT across quantiles with query correlation	196
6.4	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	197
6.5	Performance of negative correlation and SLIPT	199
6.6	Performance of simulations on a simple graph	203
6.7	Performance of simulations is similar in simple graphs	204
6.8	Performance of simulations on a graph	205
6.9	Performance of simulations on a large graph	206
6.10	Performance of simulations on a simple inhibiting graph	208
6.11	Performance is higher on a simple inhibiting graph	209
6.12	Performance of simulations on an inhibiting graph	210
6.13	Performance is affected by inhibition in graphs	211
6.14	Performance is affected by inhibition in graphs	212
6.15	Performance is affected by inhibition in graphs	213
6.16	Performance is affected by inhibition in graphs	214
6.17	Performance is affected by inhibition in graphs	215
6.18	Performance of simulations including a simple graph	217
6.19	Performance on a simple graph improves with more genes	218
6.20	Performance of simulations including a graph structure	219
6.21	Performance of simulations including an inhibiting graph	220
6.22	Performance on an inhibiting graph improves with more genes	221
6.23	Performance of simulations including a graph structure	222
6.24	Performance of simulations including an inhibiting graph	223

6.25	Performance on an inhibiting graph improves with more genes	224
6.26	Performance of simulations on the PI3K cascade	225
6.27	Performance of simulations including the PI3K cascade	226
6.28	Performance on pathways improves with more genes	228
A.1	Correlation profiles of removed samples	261
A.2	Correlation analysis and sample removal	262
A.3	Replicate excluded samples	263
A.4	Replicate samples with all remaining	264
A.5	Replicate samples with some excluded	265
D.1	Synthetic lethal expression profiles of analysed samples	282
D.2	Comparison of mtSLIPT to siRNA	284
D.3	Compare mtSLIPT and siRNA genes with correlation	288
D.4	Compare mtSLIPT and siRNA genes with correlation	288
D.5	Compare mtSLIPT and siRNA genes with siRNA viability	289
D.6	Somatic mutation locus	291
D.7	Somatic mutation against PIK3CA metagene	292
D.8	Somatic mutation against PI3K protein	293
D.9	Somatic mutation against AKT protein	294
E.1	Pathway metagene expression profiles	296
E.2	Expression profiles for constituent genes of PI3K	297
E.3	Expression profiles for p53 related genes	298
E.4	Expression profiles for estrogen receptor related genes	299
E.5	Expression profiles for BRCA related genes	300
F.1	Comparison of SLIPT in stomach to siRNA	304
G.1	Synthetic lethal expression profiles of stomach samples	313
G.2	Comparison of mtSLIPT in stomach to siRNA	315
H.1	Synthetic lethal partners across query genes	321
J.1	Synthetic Lethality in the PI3K/AKT Pathway	329
J.2	Synthetic Lethality in the PI3K/AKT Pathway in Cancer	330
J.3	Synthetic Lethality in the Extracellular Matrix	331
J.4	Synthetic Lethality in the GPCRs	332
J.5	Synthetic Lethality in the GPCR Downstream	333
J.6	Synthetic Lethality in the Translation Elongation	334
J.7	Synthetic Lethality in the Nonsense-mediated Decay	335
J.8	Synthetic Lethality in the 3' UTR	336
K.1	Synthetic Lethality and Vertex Degree	337
K.2	Synthetic Lethality and Centrality	338
K.3	Synthetic Lethality and PageRank	339
L.1	Information centrality distribution	343

M.1	Synthetic Lethality and Heirarchy Score in PI3K	344
M.2	Heirarchy Score in PI3K against Synthetic Lethality in PI3K	345
M.3	Structure of Synthetic Lethality in PI3K	345
M.4	Structure of Synthetic Lethality Resampling	346
N.1	Performance of χ^2 and SLIPT across quantiles	347
N.2	Performance of χ^2 and SLIPT across quantiles	349
N.3	Performance of χ^2 and SLIPT across quantiles with more genes	351
N.4	Performance of χ^2 and SLIPT across quantiles with query correlation	353
N.5	Performance of χ^2 and SLIPT across quantiles with query correlation	355
N.6	Performance of χ^2 and SLIPT across quantiles with query correlation and more genes	357
O.1	Simple graph structures	359
O.2	Simple graph structure	360
O.3	Constructed graph structure	360
O.4	Large constructed graph structure.	361
O.5	Branching constructed graph structure	361
O.6	Complex constructed graph structure	363
O.7	Performance of multivariate normal simulations	365
O.8	Performance of multivariate normal simulations	366
O.9	Performance of multivariate normal simulations	367
O.10	Performance of multivariate normal simulations	368
O.11	Performance of multivariate normal simulations	369
O.12	Performance of multivariate normal simulations	370
O.13	Performance of multivariate normal simulations	371
O.14	Performance of multivariate normal simulations	372
O.15	Performance of multivariate normal simulations	373
O.16	Performance of multivariate normal simulations	374
O.17	Performance of multivariate normal simulations	375
O.18	Performance of multivariate normal simulations	376
O.19	Performance of multivariate normal simulations	377
O.20	Performance is affected by inhibition in graphs	378
O.20	Performance is affected by inhibition in graphs	379
O.21	Performance is affected by inhibition in graphs	380
O.22	Performance is affected by inhibition in graphs	381
O.23	Performance of multivariate normal simulations	383
O.24	Performance of multivariate normal simulations	384
O.25	Performance of multivariate normal simulations	385
O.26	Performance of multivariate normal simulations	386
O.27	Performance of multivariate normal simulations	388
O.28	Performance of multivariate normal simulations	389
O.29	Performance of multivariate normal simulations	390
O.30	Performance of multivariate normal simulations	391
O.31	Performance of multivariate normal simulations	392
O.32	Performance of multivariate normal simulations	393

O.33 Performance of multivariate normal simulations	394
O.34 Performance of multivariate normal simulations	395
O.35 Performance of multivariate normal simulations	396
O.36 Performance of multivariate normal simulations	397
O.37 Performance of multivariate normal simulations	398

List of Tables

1.1	Methods for Predicting Genetic Interactions	38
1.2	Methods for Predicting Synthetic Lethality in Cancer	39
1.3	Methods used by Wu <i>et al.</i> (2014)	40
2.1	Excluded Samples by Batch and Clinical Characteristics.	63
2.2	Computers used during Thesis	72
2.3	Linux Utilities and Applications used during Thesis	72
2.4	R Installations used during Thesis	73
2.5	R Packages used during Thesis	73
2.6	R Packages Developed during Thesis	75
4.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT	116
4.2	Pathways for <i>CDH1</i> partners from SLIPT	118
4.3	Pathway composition for clusters of <i>CDH1</i> partners from SLIPT	122
4.4	Pathway composition for <i>CDH1</i> partners from SLIPT and siRNA screen- ing	131
4.5	Pathways for <i>CDH1</i> partners from SLIPT	135
4.6	Pathways for <i>CDH1</i> partners from SLIPT and siRNA primary screen .	136
4.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from SLIPT	145
4.8	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	147
4.9	Query synthetic lethal genes with the most SLIPT partners	154
4.10	Pathways for genes with the most SLIPT partners	155
4.11	Pathways for <i>CDH1</i> partners from SLIPT in CCLE	156
4.12	Pathways for <i>CDH1</i> partners from SLIPT in breast CCLE	158
5.1	analysis of variance (ANOVA) for Synthetic Lethality and Vertex Degree	175
5.2	ANOVA for Synthetic Lethality and Information Centrality	177
5.3	ANOVA for Synthetic Lethality and PageRank Centrality	179
5.4	ANOVA for Synthetic Lethality and PI3K Hierarchy	182
5.5	Resampling for pathway structure of synthetic lethal detection methods	186
B.1	R Packages used during Thesis	267
C.1	Comparing SLIPT genes against Secondary siRNA Screen in breast cancer	276
C.2	Comparing mtSLIPT genes against Secondary siRNA Screen in breast cancer	277
C.3	Comparing SLIPT genes against Secondary siRNA Screen in stomach cancer	277

D.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from mtSLIPT . . .	279
D.2	Pathways for <i>CDH1</i> partners from mtSLIPT	280
D.3	Pathway composition for clusters of <i>CDH1</i> partners from mtSLIPT . .	283
D.4	Pathway composition for <i>CDH1</i> partners from mtSLIPT and siRNA . .	285
D.5	Pathways for <i>CDH1</i> partners from mtSLIPT	286
D.6	Pathways for <i>CDH1</i> partners from mtSLIPT and siRNA primary screen	287
D.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from mtSLIPT . .	290
F.1	Synthetic lethal gene partners of <i>CDH1</i> from SLIPT in stomach cancer	302
F.2	Pathway composition for clusters of <i>CDH1</i> partners in stomach SLIPT	303
F.3	Pathway composition for <i>CDH1</i> partners from SLIPT and siRNA screen- ing	305
F.4	Pathways for <i>CDH1</i> partners from SLIPT in stomach cancer	306
F.5	Pathways for <i>CDH1</i> partners from SLIPT in stomach and siRNA screen	307
F.6	Candidate synthetic lethal metagenes against <i>CDH1</i> from SLIPT in stomach cancer	308
G.1	Synthetic lethal gene partners of <i>CDH1</i> from mtSLIPT in stomach cancer	310
G.2	Pathways for <i>CDH1</i> partners from mtSLIPT in stomach cancer	311
G.3	Pathway composition for clusters of <i>CDH1</i> partners in stomach mtSLIPT	314
G.4	Pathway composition for <i>CDH1</i> partners from mtSLIPT and siRNA . .	316
G.5	Pathways for <i>CDH1</i> partners from mtSLIPT in stomach cancer	317
G.6	Pathways for <i>CDH1</i> partners from mtSLIPT in stomach and siRNA screen	318
G.7	Candidate synthetic lethal metagenes against <i>CDH1</i> from mtSLIPT in stomach cancer	319
H.1	Query synthetic lethal genes with the most SLIPT partners	322
H.2	Pathways for genes with the most SLIPT partners	323
I.1	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT in CCLE	325
I.2	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT in breast CCLE	326
I.3	Candidate synthetic lethal gene partners of <i>CDH1</i> from SLIPT in stom- ach CCLE	327
I.4	Pathways for <i>CDH1</i> partners from SLIPT in stomach CCLE	328
I.5	Pathways for <i>CDH1</i> partners from SLIPT in breast and stomach CCLE	328
K.1	ANOVA for Synthetic Lethality and Vertex Degree	340
K.2	ANOVA for Synthetic Lethality and Information Centrality	340
K.3	ANOVA for Synthetic Lethality and PageRank Centrality	340
L.1	Information centrality for genes and molecules in the Reactome network	342
M.1	ANOVA for Synthetic Lethality and PI3K Hierarchy	344
M.2	Resampling for pathway structure of synthetic lethal detection methods	346

Glossary

synthetic lethal Genetic interactions where inactivation of multiple genes is inviable (or deleterious) when they are viable if inactivated separately.

Acronyms

ANOVA Analysis of Variance.

SLIPT Synthetic lethal interaction prediction tool.

Chapter 6

Simulation and Modeling of Synthetic Lethal Pathways

Simulation and modelling of synthetic lethality in gene expression will be revisited in greater detail in this chapter, building upon the results provided to support the use of Synthetic Lethal Interaction Prediction Tool (SLIPT) in Section 3.3. 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 a computational approach for detecting synthetic lethal candidates from expression data throughout this thesis (in Chapters 4 and 5).

While this basic framework was sufficient to support the use of SLIPT in prior Chapters, 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 upon more complex correlation structures. Together these simulation investigations assess the performance of the SLIPT methodology, including on pathway graph structures (such as those discussed in Chapter 5) and determine 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 equivalent χ^2 test alone) were applied across a range of pa-

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

A refined simulation procedure was developed specifically to extend the simulation procedure (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 (as discussed in Chapter 5). 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 were applied in a computational pipeline across many parameters with high-performance computing (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 methods

The SLIPT methodology (as it has been applied throughout Chapters 4 and 5) was compared to alternative computational approaches to detecting synthetic lethality in simulated gene expression data. As discussed in Section 3.3, this procedure enables testing the performance of detecting 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’s correlation, the χ^2 test, and testing for bimodality) to evaluate the strengths of the SLIPT methodology, including modifications to the parameters of SLIPT.

Further testing of the performance of the SLIPT software R package (which is publicly released on GitHub as described in Section 3.5) has been left to third party researchers to impartially compare it to other software for synthetic lethal detection which is outside the scope of this thesis. The following comparisons of simulations of computational detection of synthetic lethality with different statistical rationales suffice to discuss the strengths of SLIPT, evaluate whether it is appropriate for further application in genomics research, and identify limitations which may be addressed

with further developments. Some potential avenues for further development of computational synthetic lethal discovery will be discussed in Section 7.2.

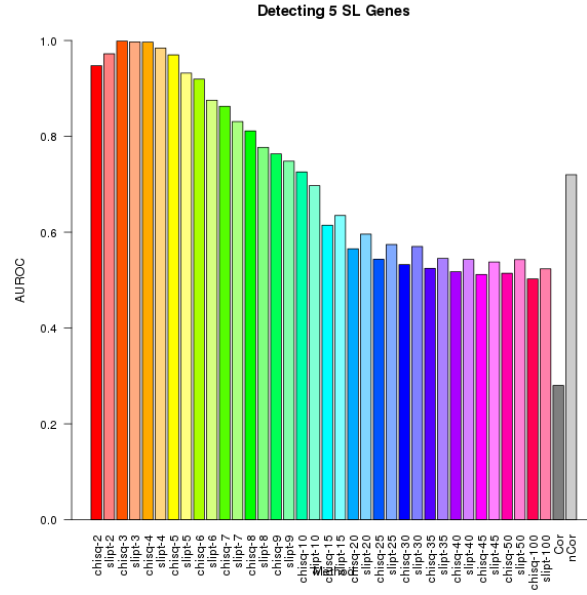
6.1.1 Performance of SLIPT and χ^2 across Quantiles

Simulated datasets with synthetic lethal partner genes were generated using the multivariate normal simulation procedure (as described in Section 3.2.2) with performance assessed using area under the receiver operating characteristic (AUROC) analysis (as described in Section 2.3.5). Synthetic lethal detection was compared for modifications to the SLIPT methodology (as described in Section 3.1), namely that the quantiles used to define low and high expression was varied. Rather than $\frac{1}{3}$ (as used throughout this thesis) the samples below the lowest $\frac{1}{n}$ quantile and above the highest $\frac{1}{n}$ 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 $\frac{1}{2}$ quantile (the median), to 100, using the lowest (1%) and highest (99%) percentiles.

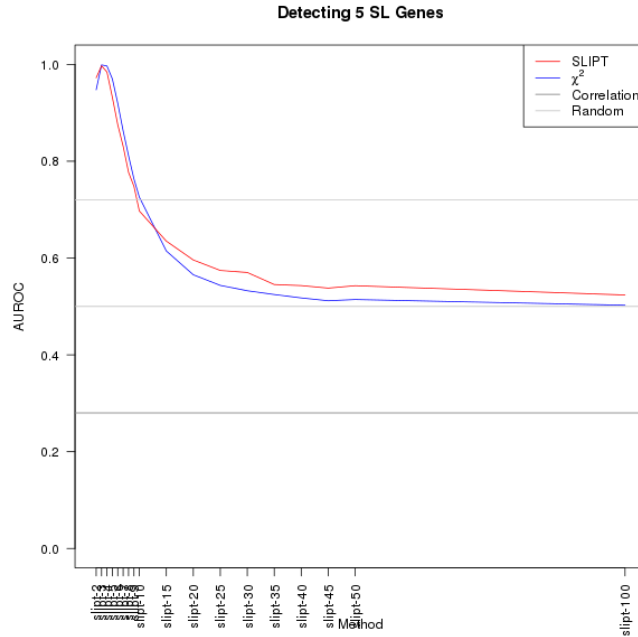
This enables testing of the threshold for lowly expressing genes which is most able to distinguish synthetic lethal genes, even with higher-order synthetic lethal interactions (as discussed in Section 3.2.1). Both SLIPT with the directional criteria for synthetic lethality and significance of the equivalent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

The results presented throughout this section use the example of 5 synthetic lethal partners to illustrate the differences in performance between the standard SLIPT procedure (slipt-3) to n quantiles (slipt- n), the χ^2 -test on the same quantiles, and positive or negative correlation. However, similar results across different numbers of known synthetic lethal genes are shown in Appendix N. 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's 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 correla-

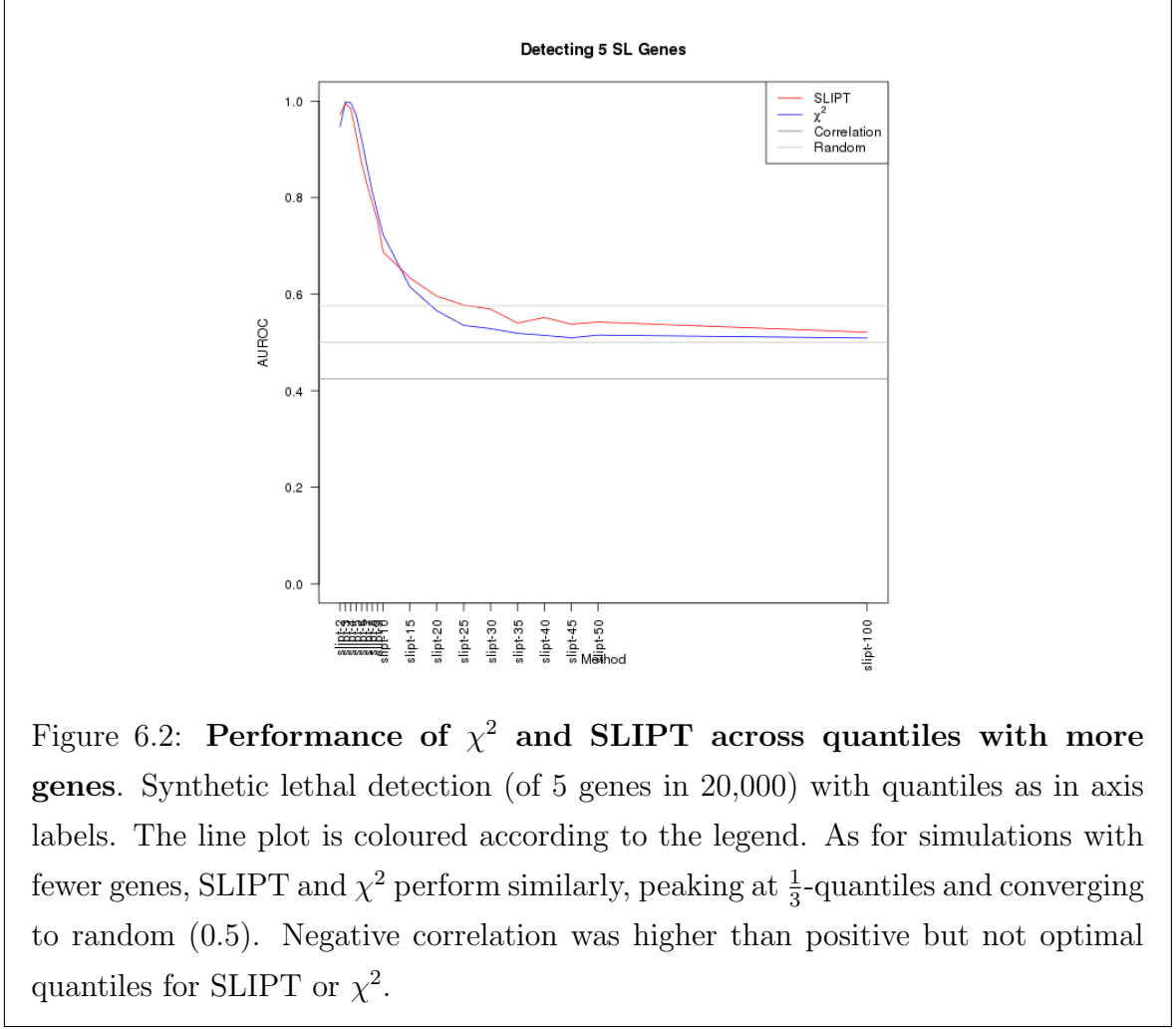


(a) Barplot of χ^2 , SLIPT, and correlation.



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

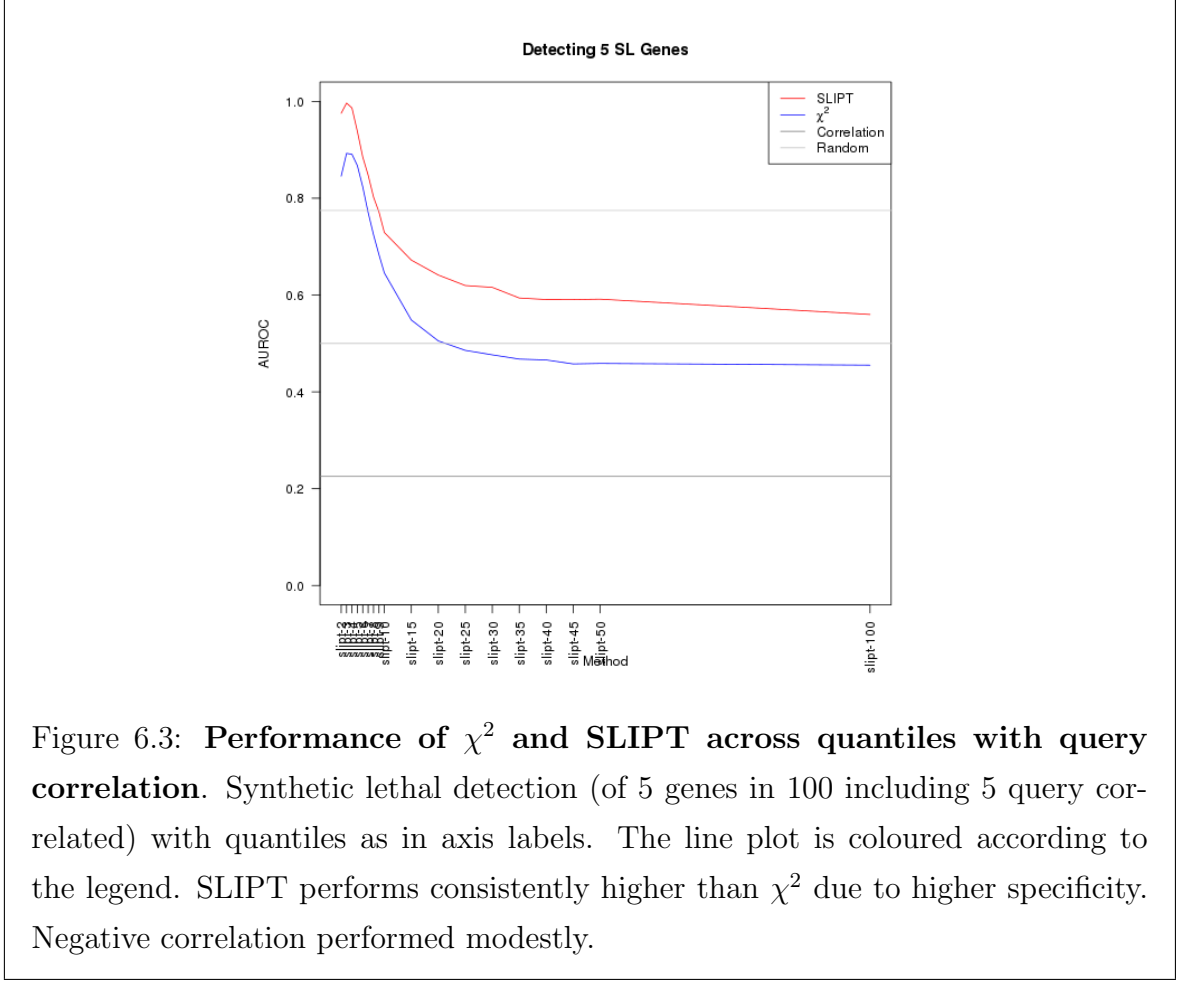
Figure 6.1: **Performance of χ^2 and SLIPT across quantiles.** Synthetic lethal detection (of 5 genes) with quantiles as in axis labels. The barplot uses the same hues for each quantile (grey for correlation) and darker for χ^2 (and positive correlation). The line plot is coloured according to the legend.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



tion 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 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 N.1 and N.2. These demonstrate that the findings shown for 5 synthetic lethal genes is 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 human of genes in a human gene expression dataset and assess the impact of a higher proportion of non synthetic lethal genes (potential



false positives). Simulated datasets of this size were also used in Section 3.3.2 to test the specificity in a number of genes similar to that in experimental datasets for cancer genomes. As shown in Figure 6.2, the above findings were replicated in simulations of a larger dataset with 20,000 genes. These were also robustly replicated across varying numbers of underlying synthetic lethal genes (as shown in Figure N.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 of on the performance of synthetic lethal detection. SLIPT able to distinguish these from synthetic lethal partners and hence has 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

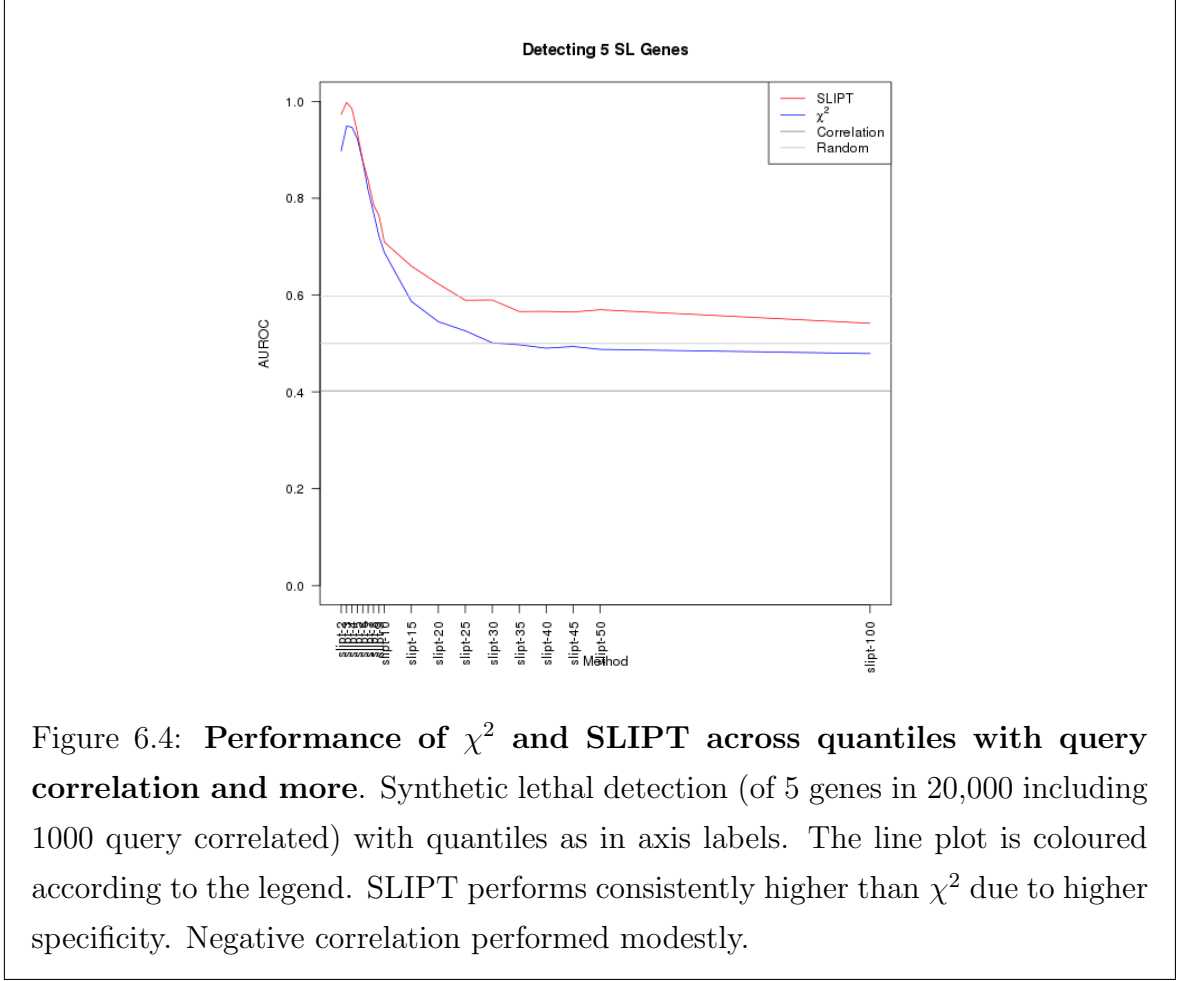


Figure 6.3, this specificity is reflected in the increased AUROC performance values for SLIPT (in contrast to Figure 6.1). This specificity can be attributed to the directional criteria (as described in Section 3.1) since the χ^2 -test alone performs comparatively poorly with positively correlated genes.

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 feasible in 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. However, the difference between SLIPT and χ^2 were less pronounced in a larger dataset since the sheer number of negative genes (as potential false positives) affects the specificity of SLIPT which distinguishes it from χ^2 -test alone and is an important consideration in large-scale genomics analysis.

Nevertheless, SLIPT with 3-quantiles (as performed throughout Chapters 4 and 5), has higher performance than other quantiles, particularly with positive correlations (replicating the Section 3.3.2.2). These findings hold across different numbers of underlying synthetic lethal genes (as shown in Figures N.5 and N.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 was sufficient sample size when the data is divided into quantiles. The SLIPT methodology further performs better for 3-quantiles (and other moderate values), irrespective of sample size or p-value threshold as AUROC values are independent from them. Such quantiles ensure 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 as both significance and the SLIPT directional conditions rely use the same expected values.

6.1.2 Alternative Synthetic Lethal Detection Strategies

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

The BImodal Subsetting ExPression (BiSEp) R package (Wappett, 2014) for using bimodality to detect synthetic lethality (Wappett *et al.*, 2016) and linear models were also considered. These statistical methods span a range of computational approaches to detecting synthetic lethality and serve to compare alternatives to SLIPT, supporting its design (see Section 3.1) and application (in Chapters 4 and 5). Although these are intended not intended to be a comprehensive benchmarking of existing synthetic lethal tools, implementing other synthetic lethal detection software is out of the scope of this project. 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's) 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. Although negative correlation still often performed considerably better than random chance.

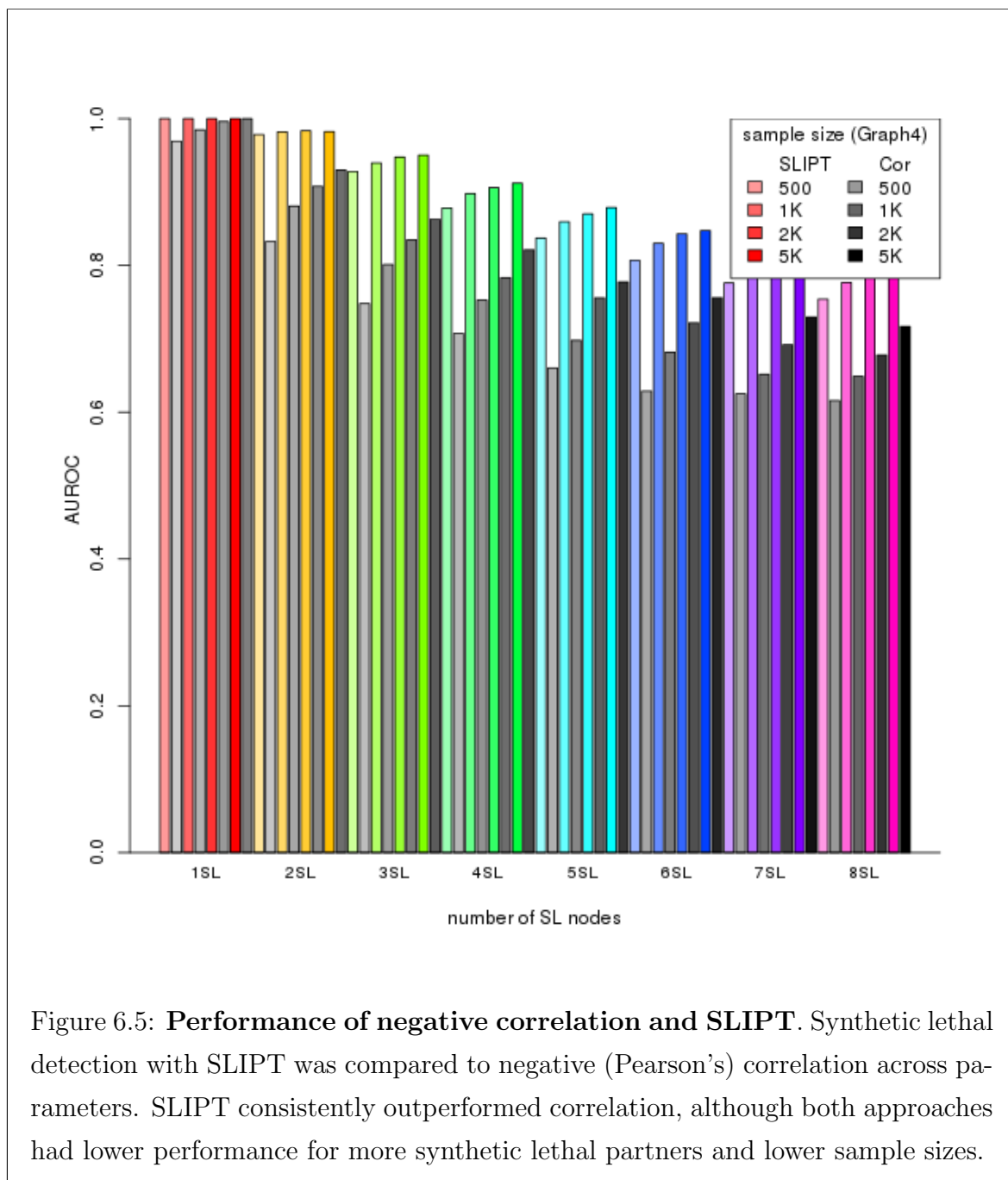


Figure 6.5: **Performance of negative correlation and SLIPT.** Synthetic lethal detection with SLIPT was compared to negative (Pearson’s) correlation across parameters. SLIPT consistently outperformed correlation, although both approaches had lower performance for more synthetic lethal partners and lower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

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 (as explored in more detail in Section 6.2). 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.

[Add Other Graphs to Appendix?]

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

6.1.2.2 Testing for Bimodality with BiSEp

Exhaustive 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 were 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 simulations. 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 are sufficient to evaluate SLIPT and compare it to other statistical rationales. An comprehensive comparison to contemporary synthetic lethal detection approaches (and those released in the future) or further benchmarking is left to an impartial researcher to evaluate and is outside the scope of this thesis. 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.

[Discuss linear models?]

6.2 Simulations with Graph Structures

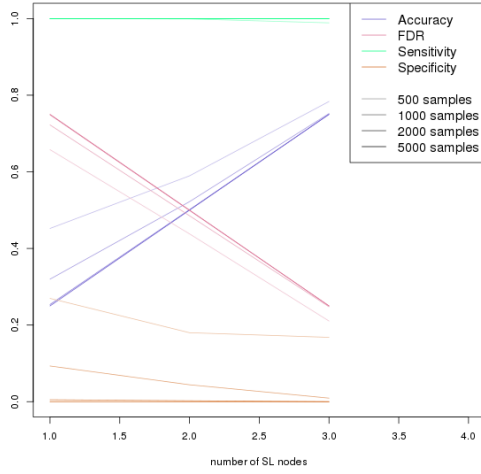
Simulations of synthetic lethality in Section 3.3 included correlated blocks of genes as a rudimentary model of pathway structure and co-regulated genes. Here the simulation procedure was expanded to account for more complex graph structures by sampling from multivariate normal distributions with correlation 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 (such as 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 and on 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 affects synthetic lethal detection, and whether synthetic lethal detection varies across which gene is synthetic lethal or affects proximal genes in 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.

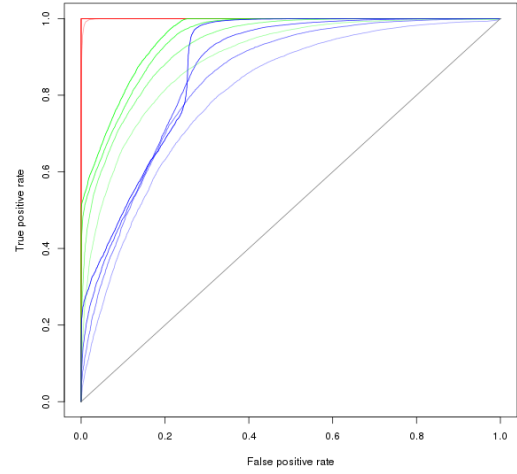
6.2.1 Performance over a Graph Structure

6.2.1.1 Simple Graph Structures

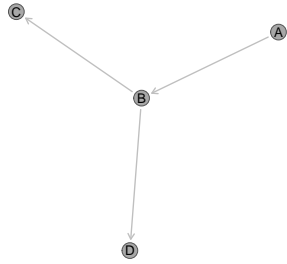
6.2.1.2 Constructed Graph Structures



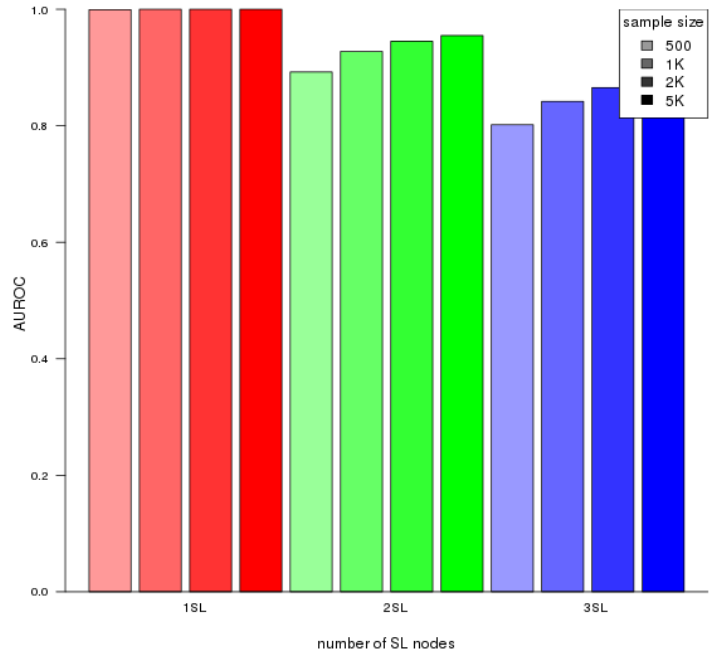
(a) Statistical evaluation



(b) Receiver operating characteristic

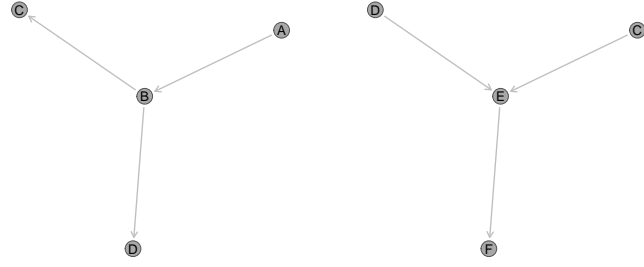


(c) Graph Structure



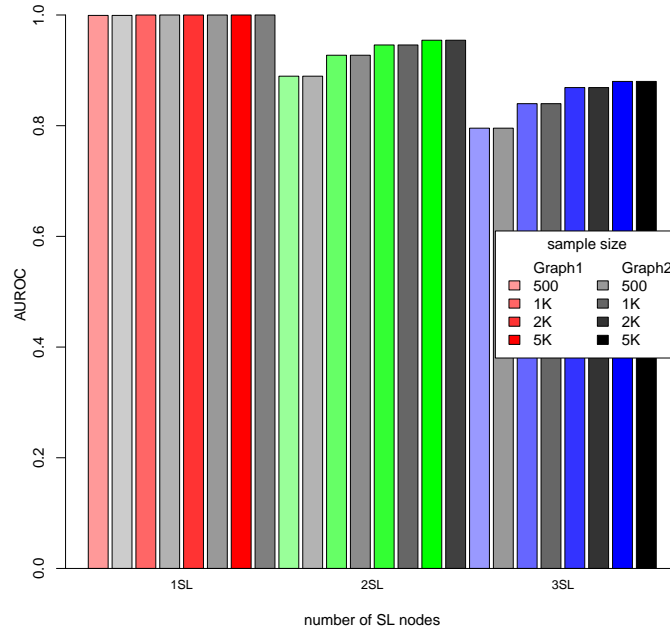
(d) Statistical performance

Figure 6.6: Performance of simulations on a simple graph. 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.



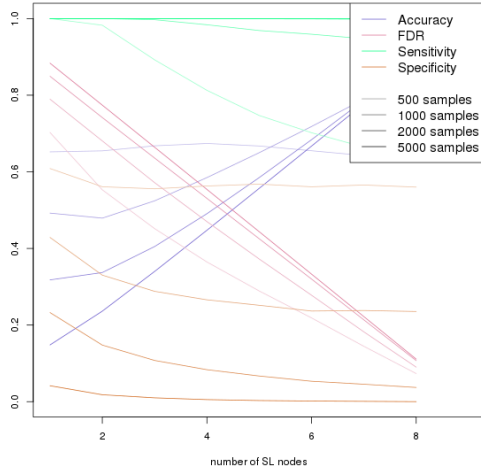
(a) Graph Structure

(b) Graph Structure

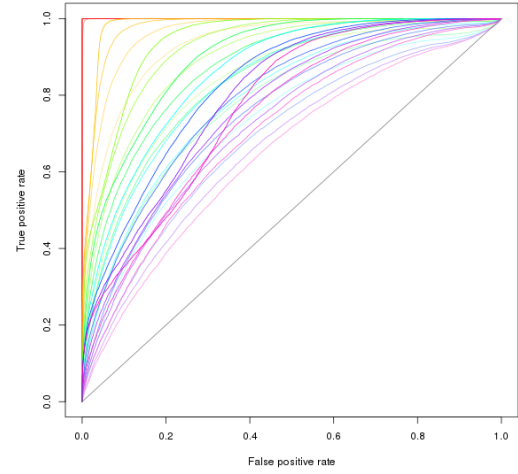


(c) Gene category in simulations

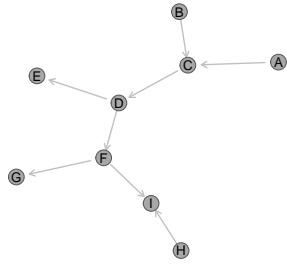
Figure 6.7: Performance of simulations is similar in simple graphs. The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations. 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.



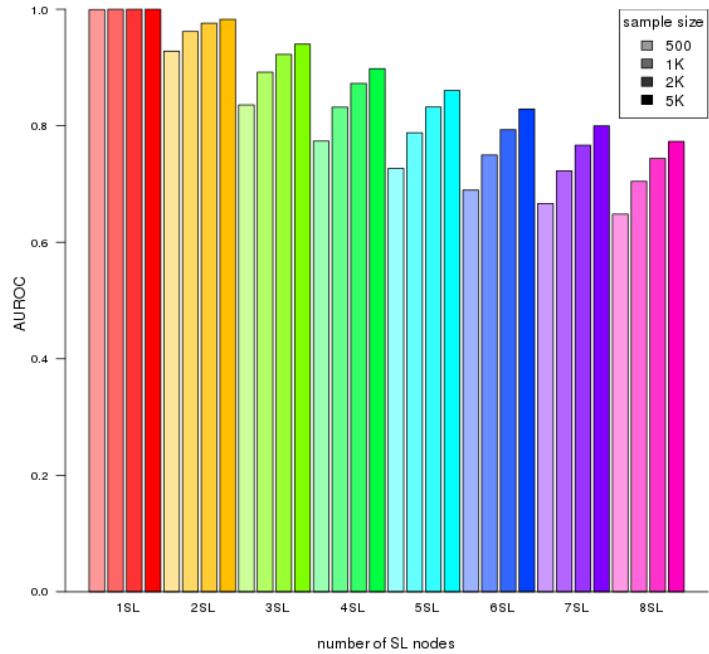
(a) Statistical evaluation



(b) Receiver operating characteristic

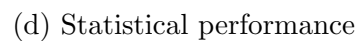
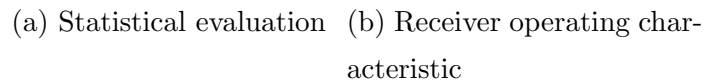


(c) Graph Structure



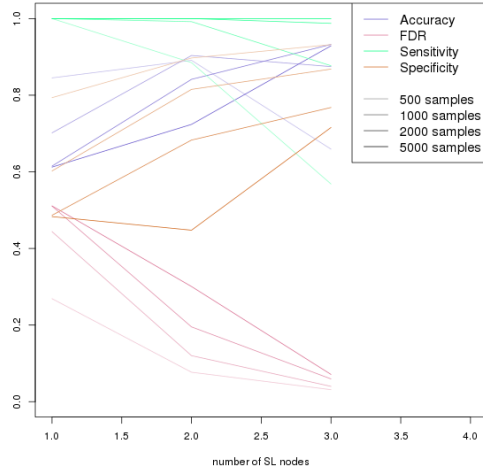
(d) Statistical performance

Figure 6.8: Performance of simulations on a graph. 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.

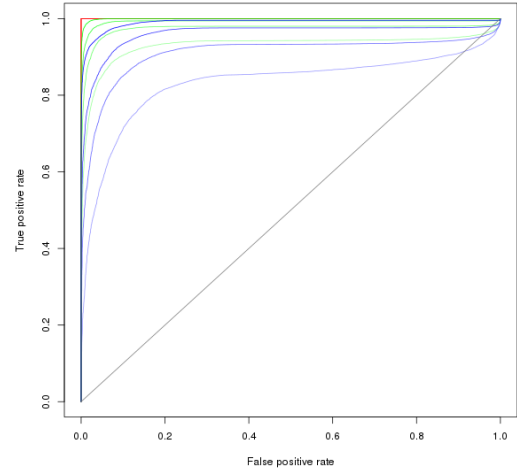


206

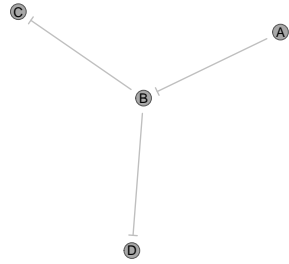
6.2.2 Performance with Inhibitions



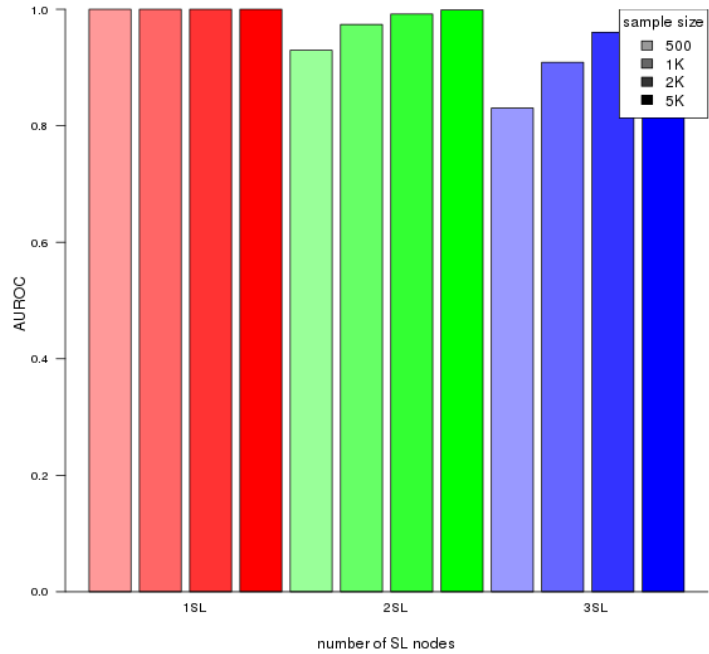
(a) Statistical evaluation



(b) Receiver operating characteristic

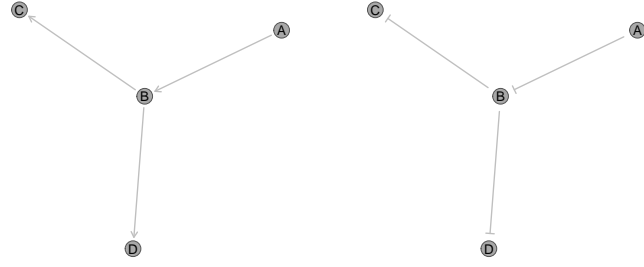


(c) Graph Structure



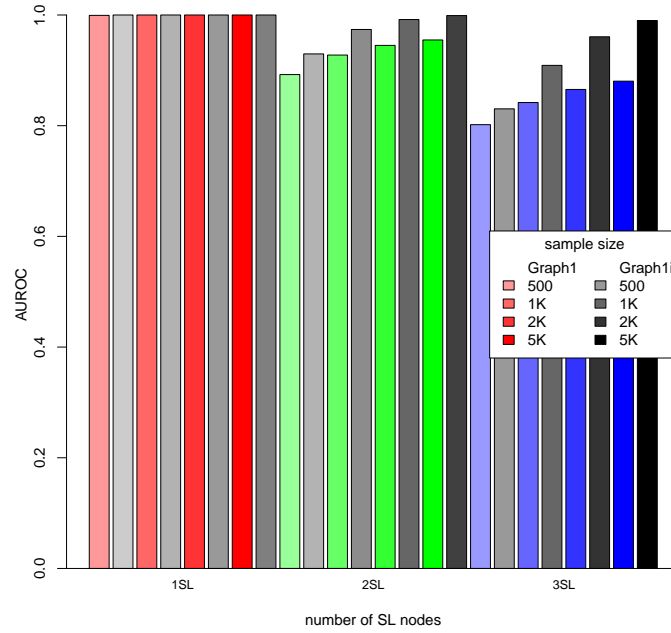
(d) Statistical performance

Figure 6.10: Performance of simulations on a simple inhibiting graph. 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.



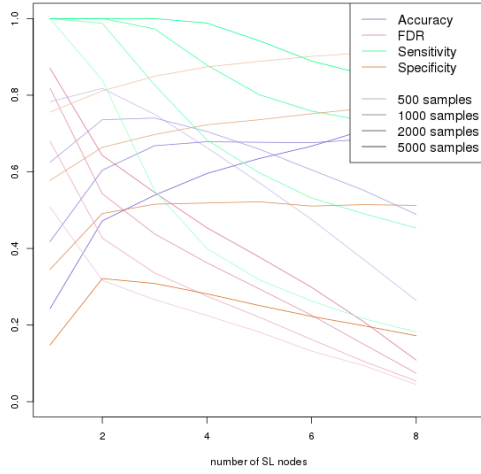
(a) Graph Structure

(b) Graph Structure

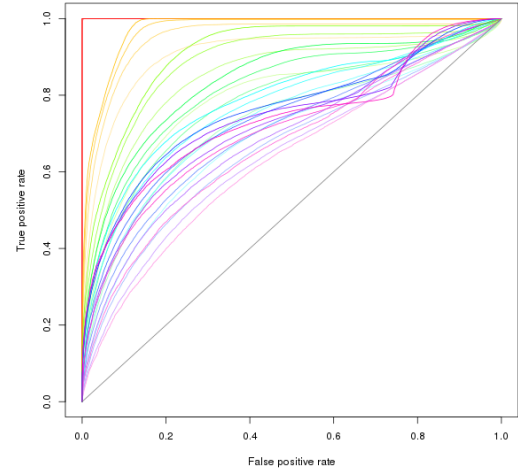


(c) Gene category in simulations

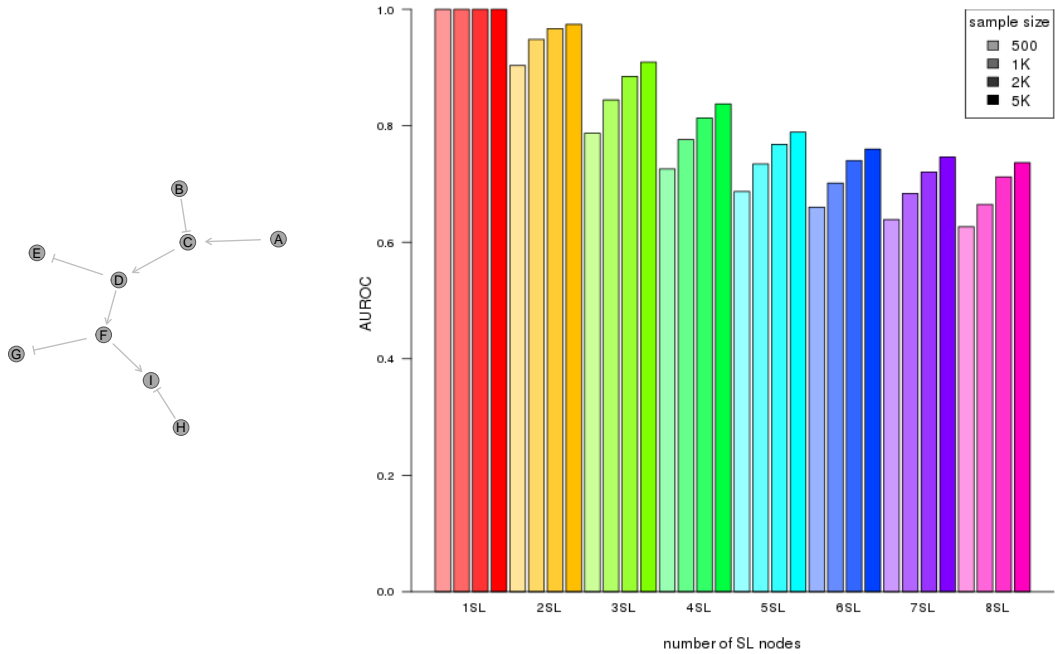
Figure 6.11: **Performance is higher on a simple inhibiting graph.** The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations. 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.



(a) Statistical evaluation



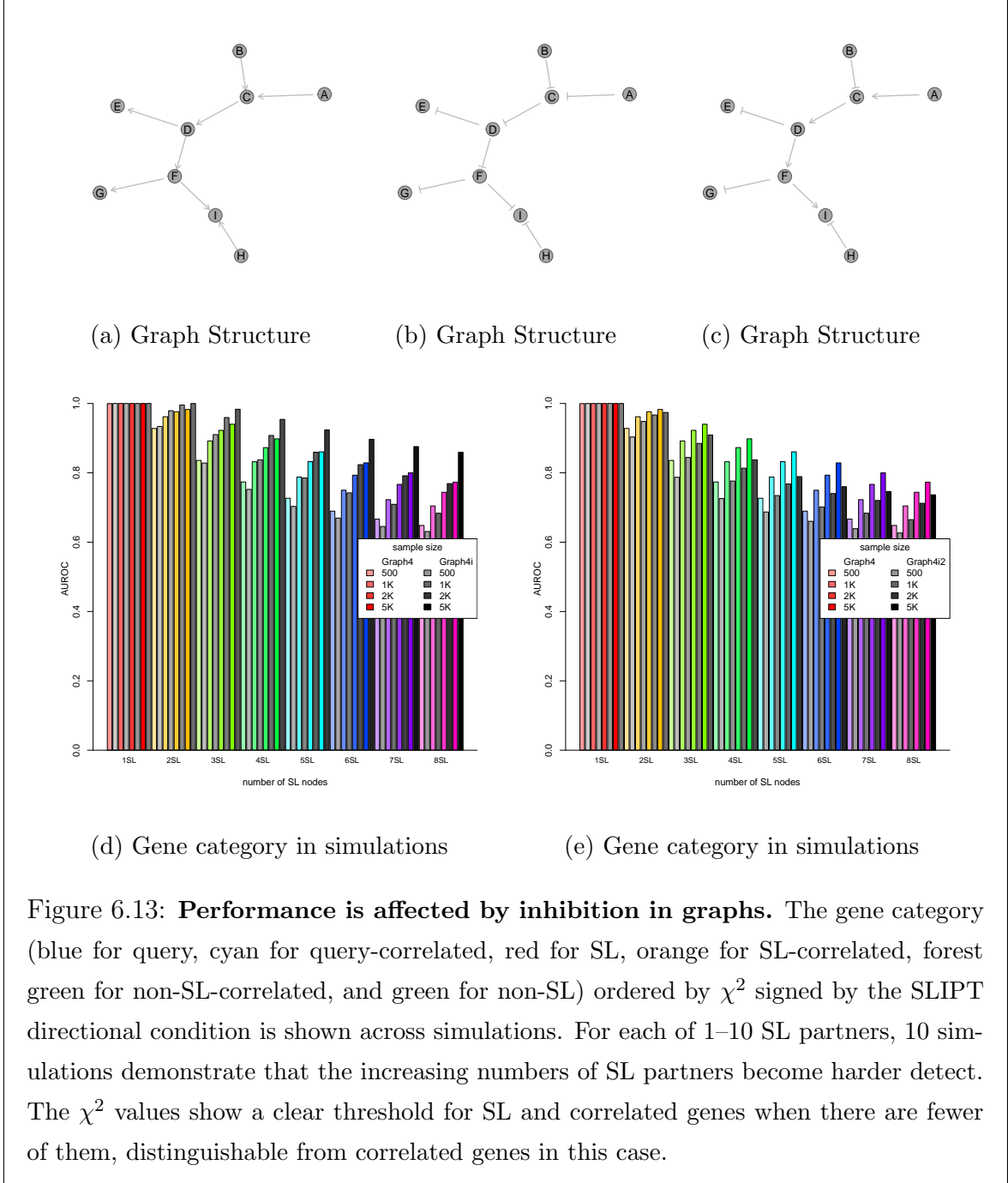
(b) Receiver operating characteristic



(c) Graph Structure

(d) Statistical performance

Figure 6.12: Performance of simulations on an inhibiting graph. 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.



6.2.3 Synthetic Lethality across Graph Structures

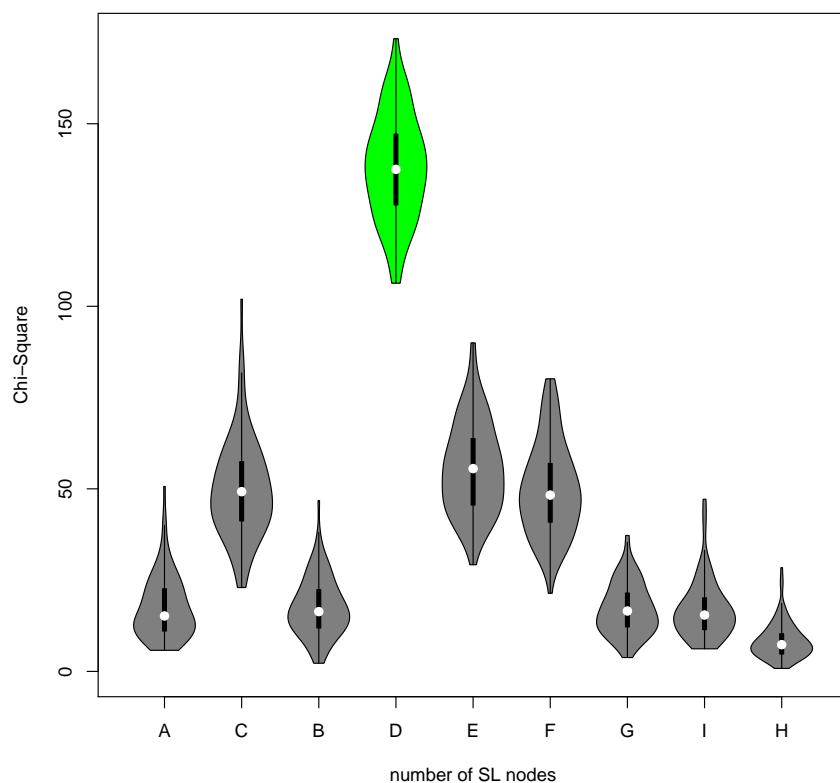
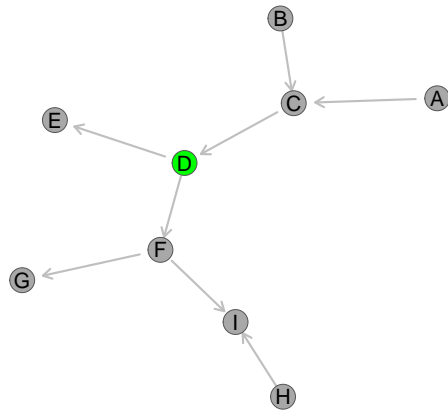
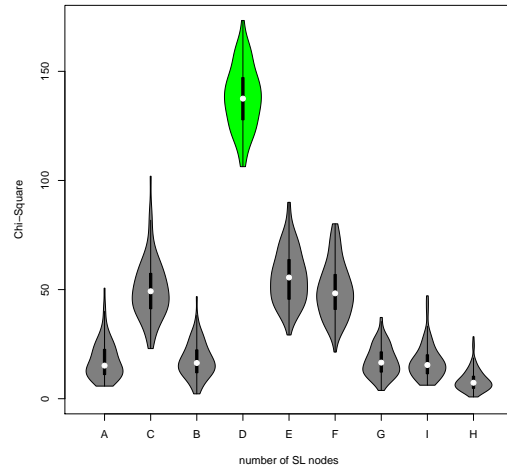


Figure 6.14: **Performance is affected by inhibition in graphs.** The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations. 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.

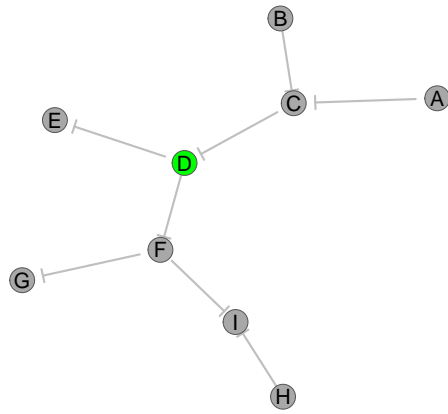


(a) Graph Structure

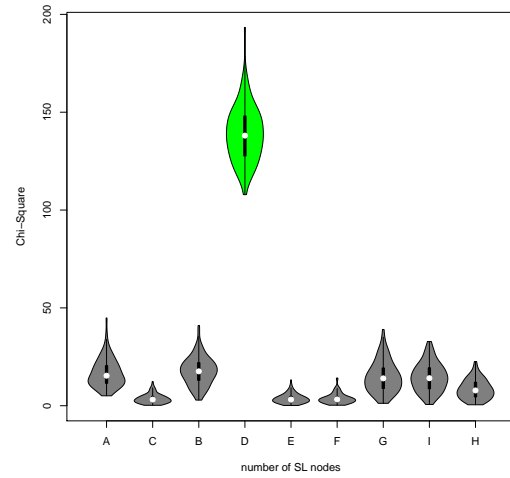


(b) Graph Structure

Figure 6.15: **Performance is affected by inhibition in graphs.** 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.

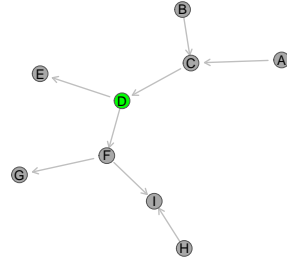


(a) Graph Structure

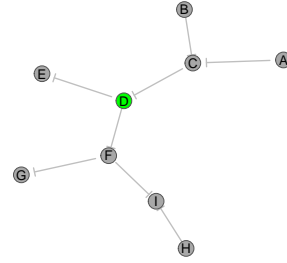


(b) Graph Structure

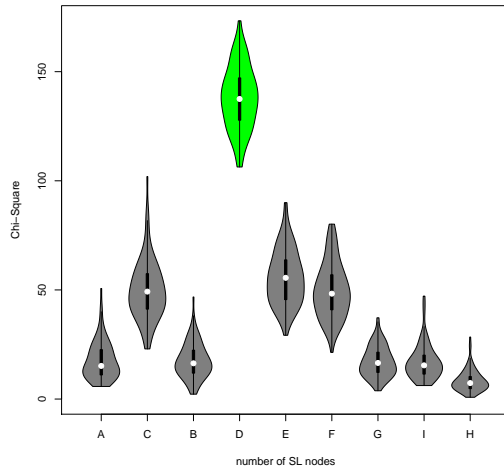
Figure 6.16: **Performance is affected by inhibition in graphs.** The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations. 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.



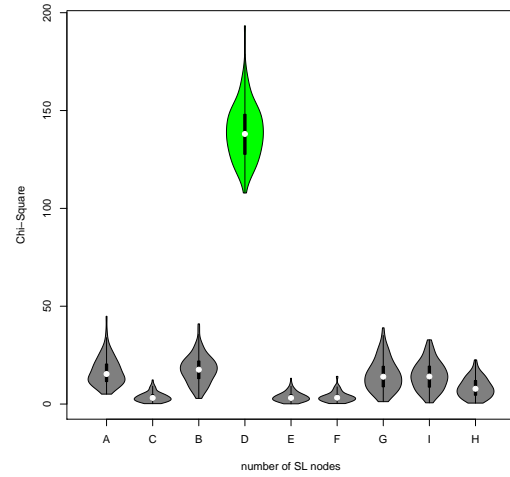
(a) Graph Structure



(b) Graph Structure



(c) Graph Structure



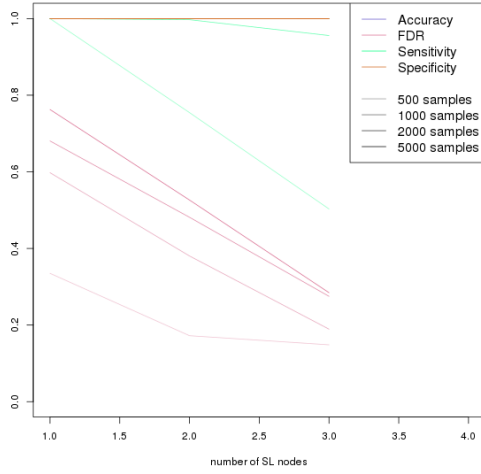
(d) Graph Structure

Figure 6.17: Performance is affected by inhibition in graphs. 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.

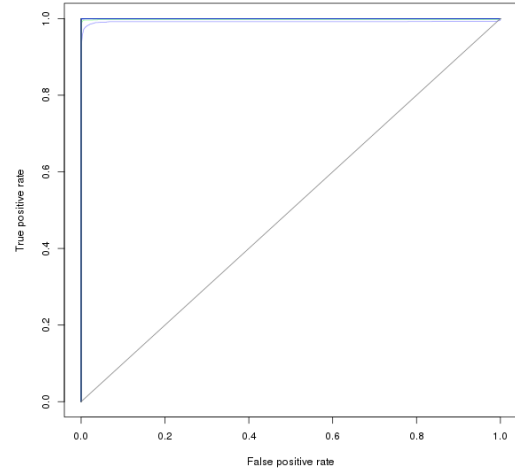
6.2.4 Performance with Feasible Gene Numbers (20,000)

6.2.4.1 Simple Graph Structures in a Genome

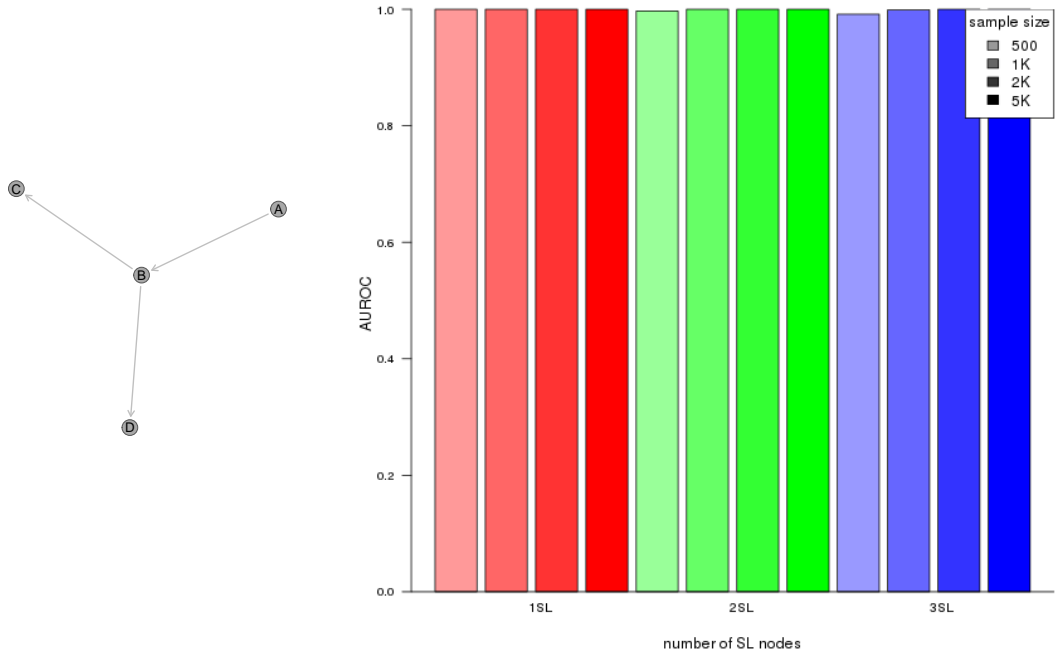
6.2.4.2 Constructed Graph Structures in a Genome



(a) Statistical evaluation



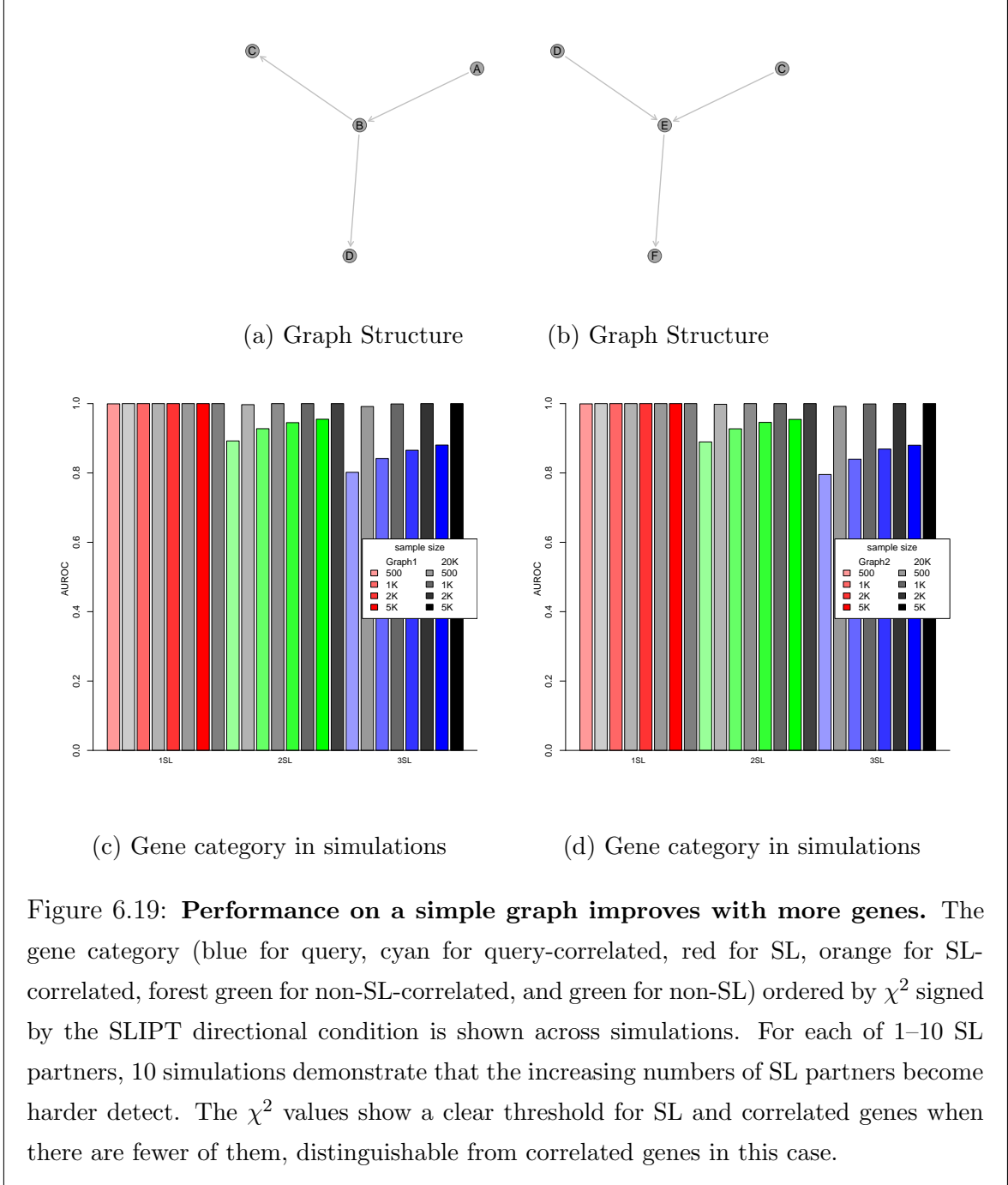
(b) Receiver operating characteristic

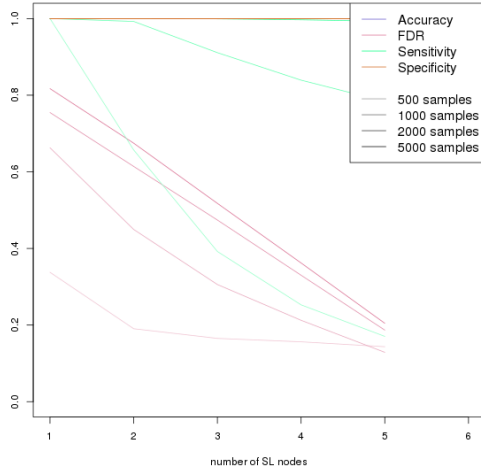


(c) Graph Structure

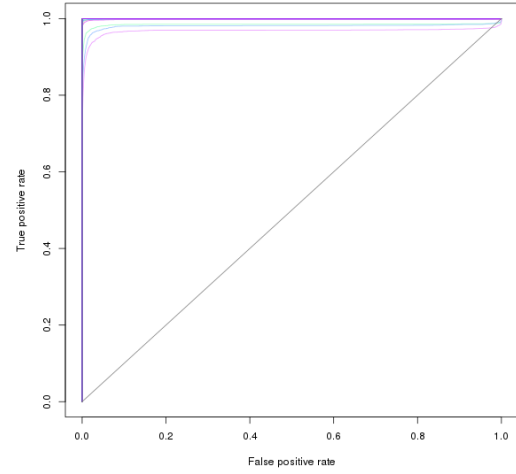
(d) Statistical performance

Figure 6.18: Performance of simulations including a simple graph. 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.

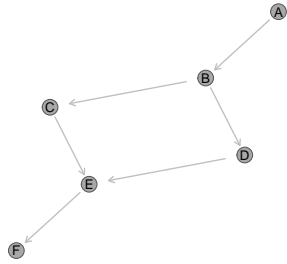




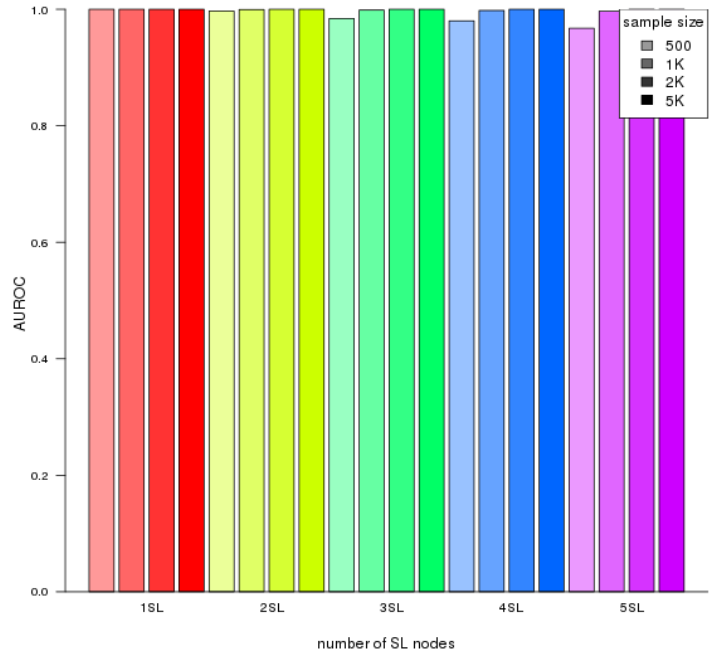
(a) Statistical evaluation



(b) Receiver operating characteristic

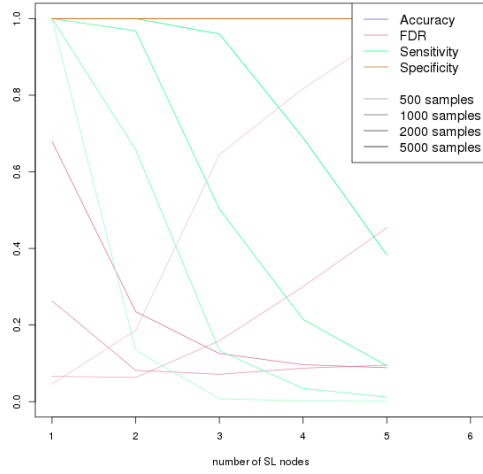


(c) Graph Structure

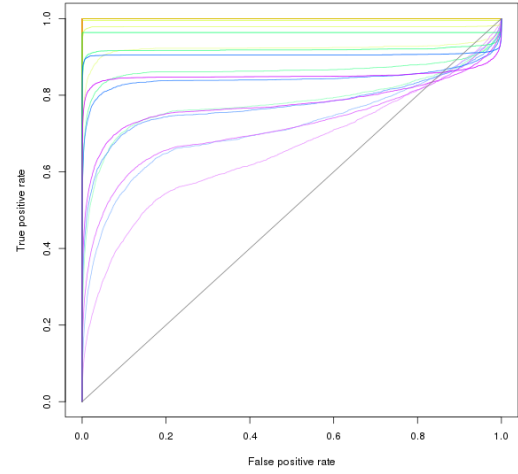


(d) Statistical performance

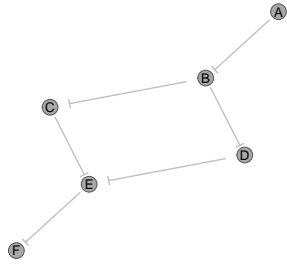
Figure 6.20: Performance of simulations including a graph structure. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (without correlation structure). Performance of SLIPT declines for more synthetic partners but this is mitigated by increased sample sizes (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.



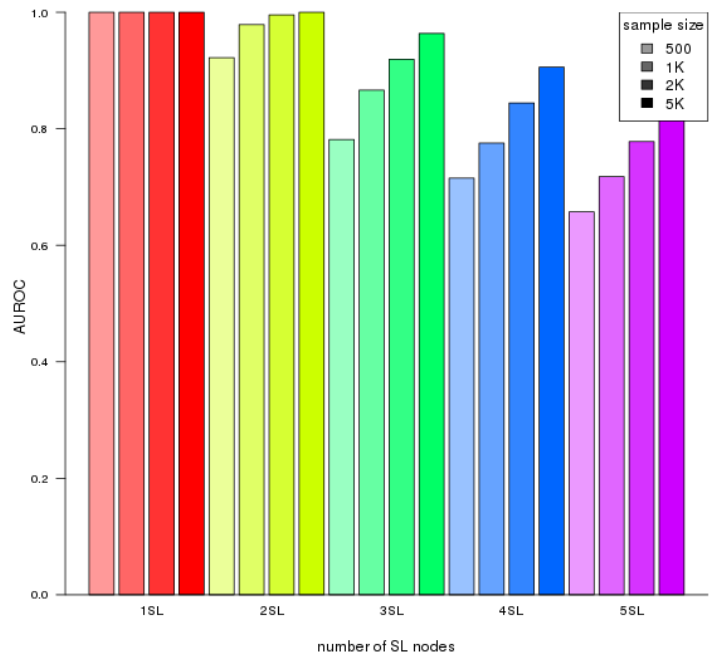
(a) Statistical evaluation



(b) Receiver operating characteristic

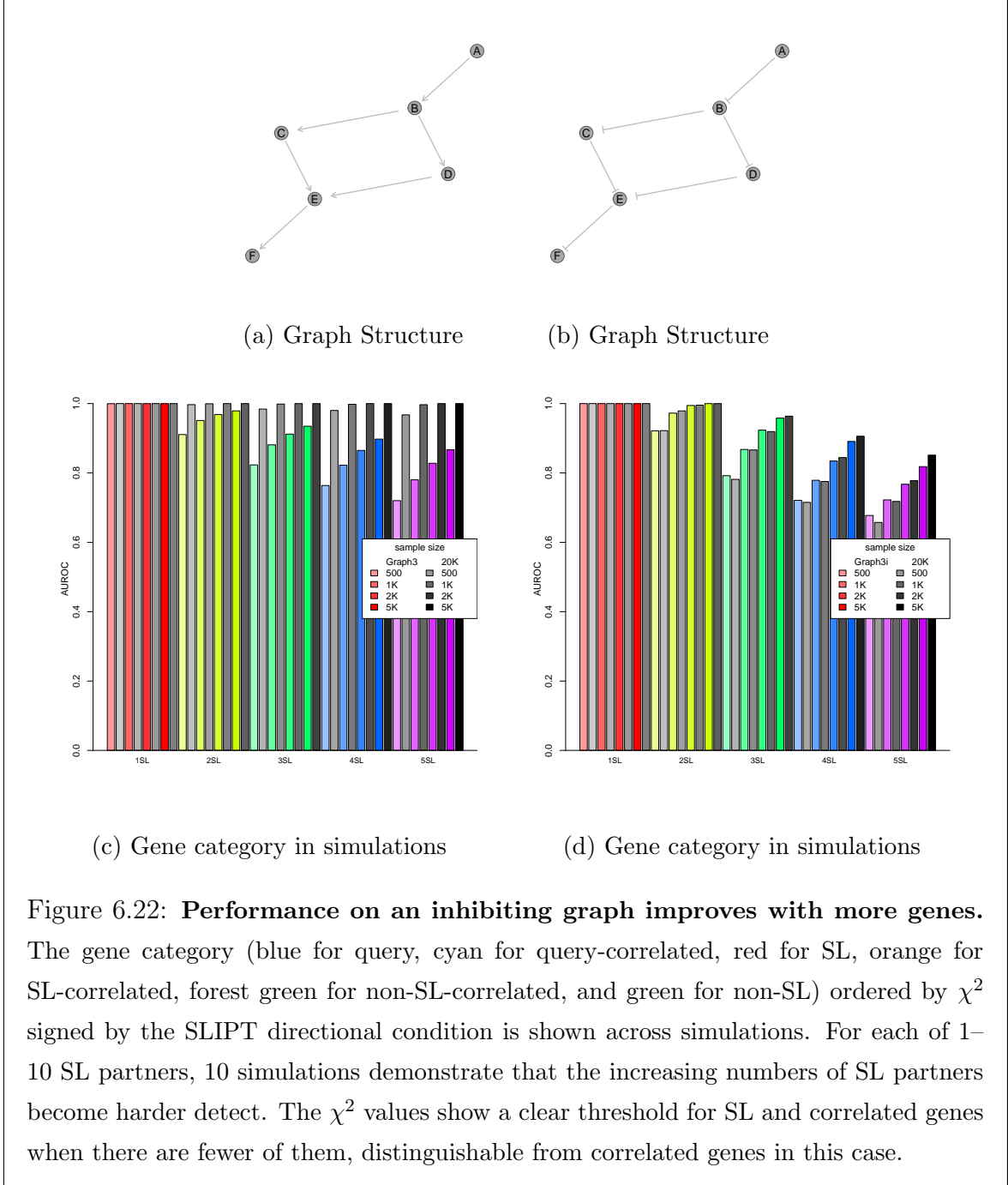


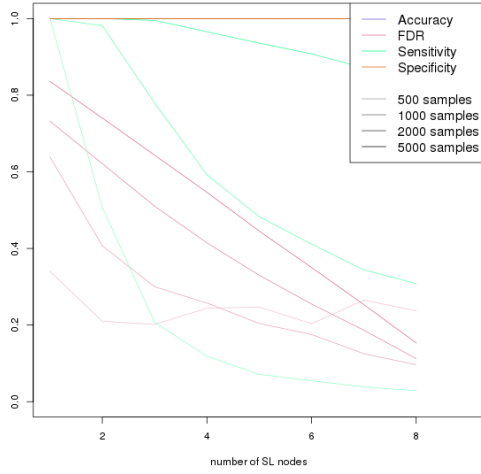
(c) Graph Structure



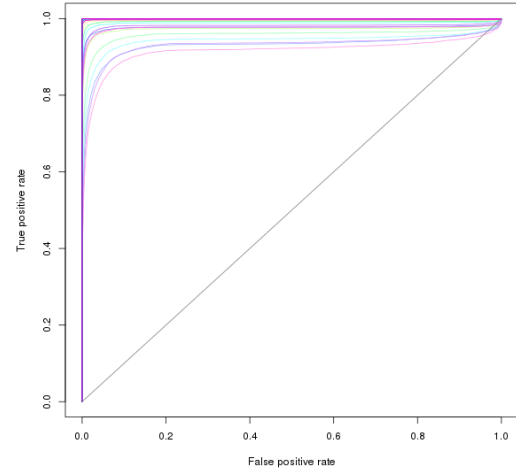
(d) Statistical performance

Figure 6.21: Performance of simulations including an inhibiting graph. 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.

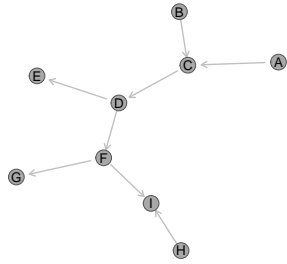




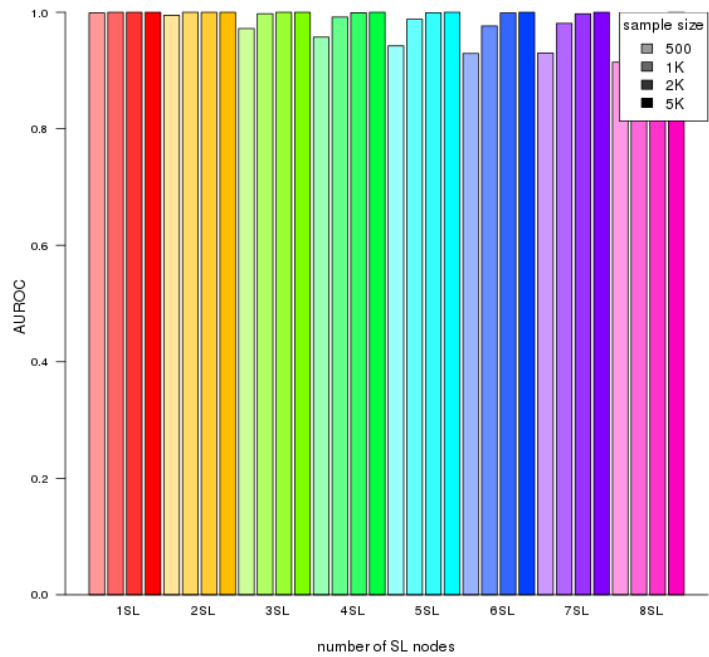
(a) Statistical evaluation



(b) Receiver operating characteristic

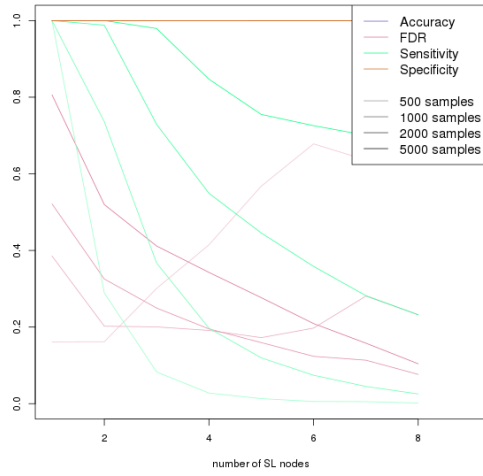


(c) Graph Structure

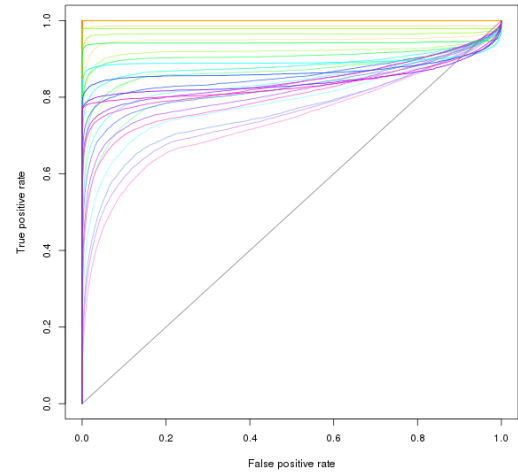


(d) Statistical performance

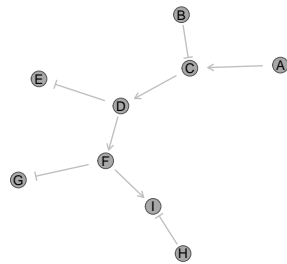
Figure 6.23: Performance of simulations including a graph structure. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (without correlation structure). Performance of SLIPT declines for more synthetic partners but this is mitigated by increased sample sizes (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.



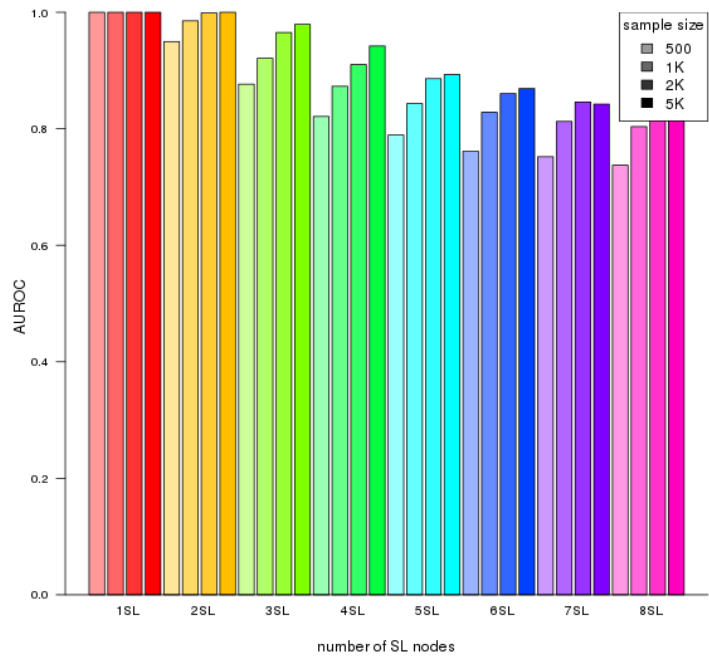
(a) Statistical evaluation



(b) Receiver operating characteristic

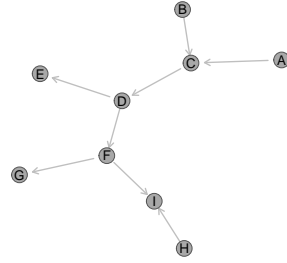


(c) Graph Structure

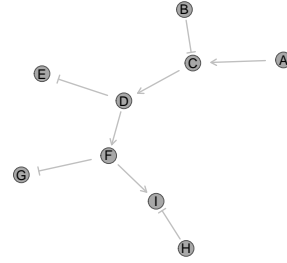


(d) Statistical performance

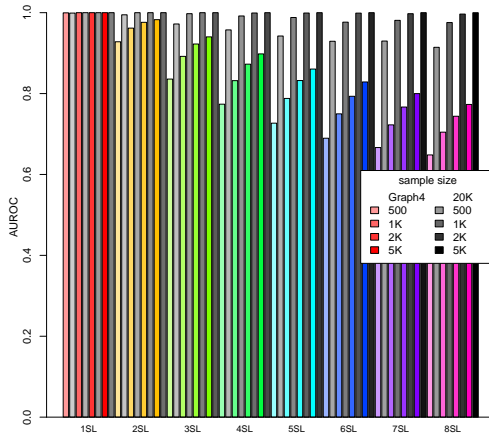
Figure 6.24: Performance of simulations including an inhibiting graph. Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (without correlation structure). Performance of SLIPT declines for more synthetic partners but this is mitigated by increased sample 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.



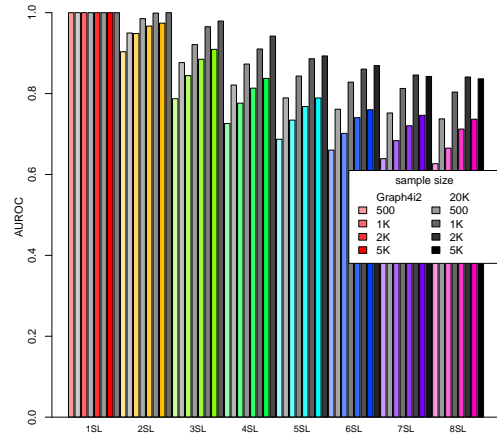
(a) Graph Structure



(b) Graph Structure



(c) Gene category in simulations



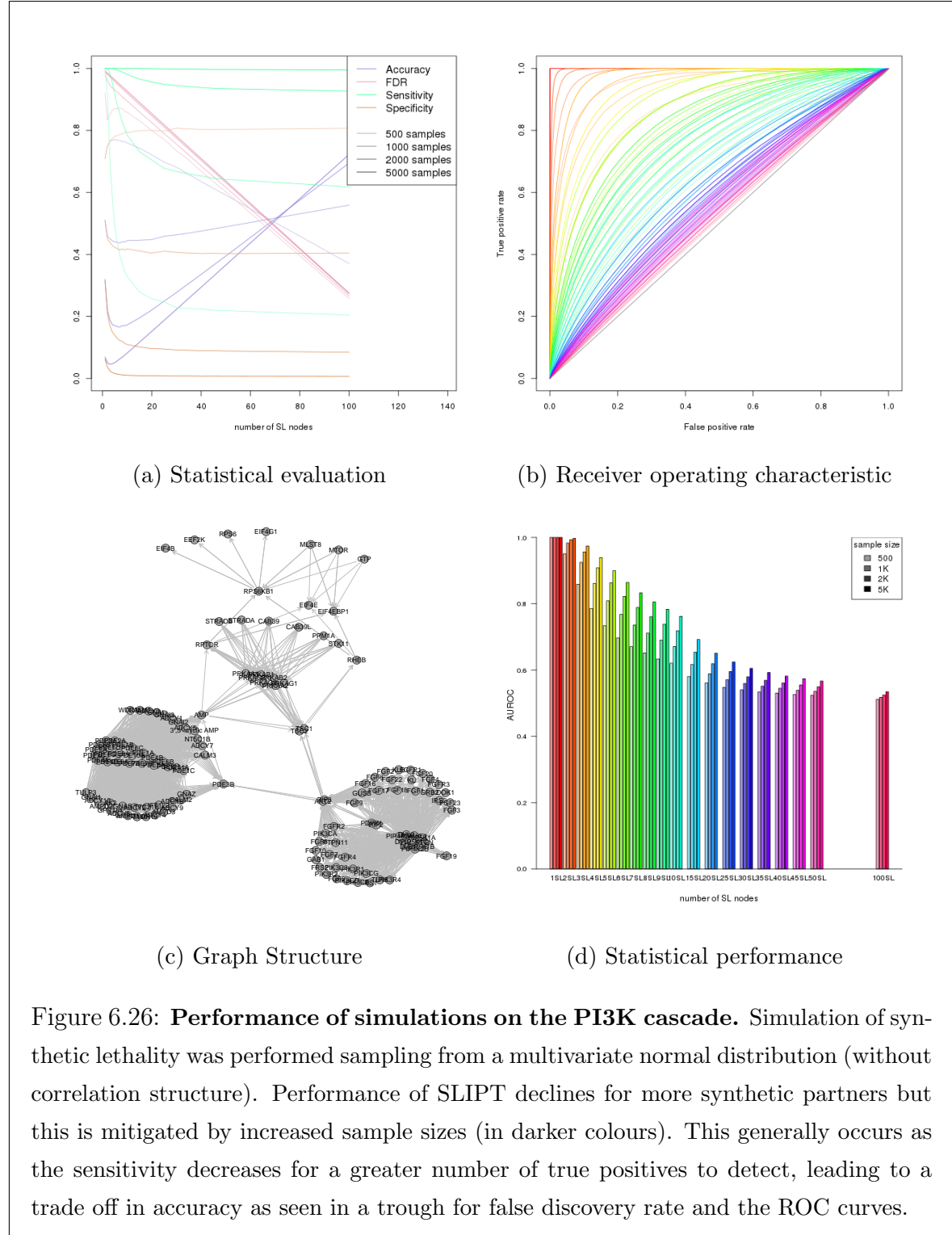
(d) Gene category in simulations

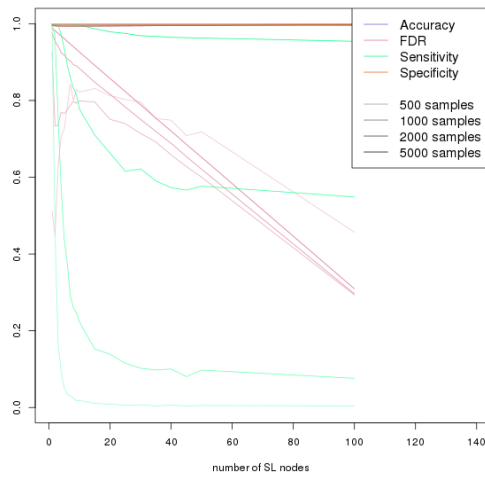
Figure 6.25: **Performance on an inhibiting graph improves with more genes.**

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.

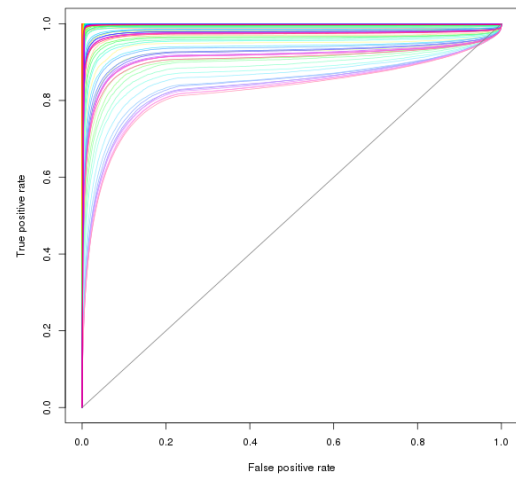
6.3 Simulations over pathway-based graphs

Text

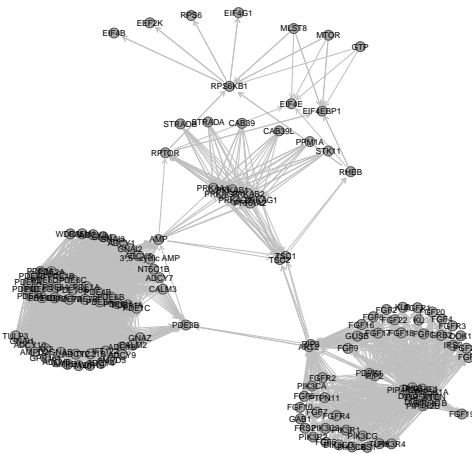




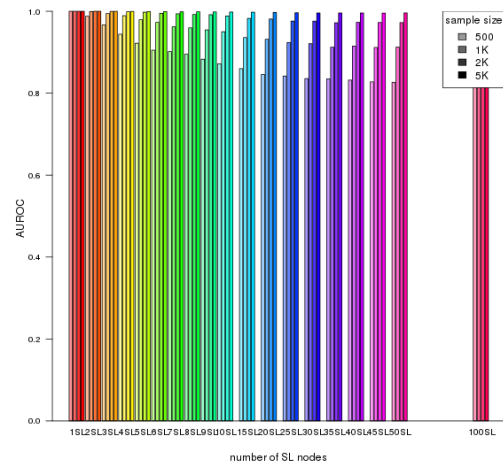
(a) Statistical evaluation



(b) Receiver operating characteristic



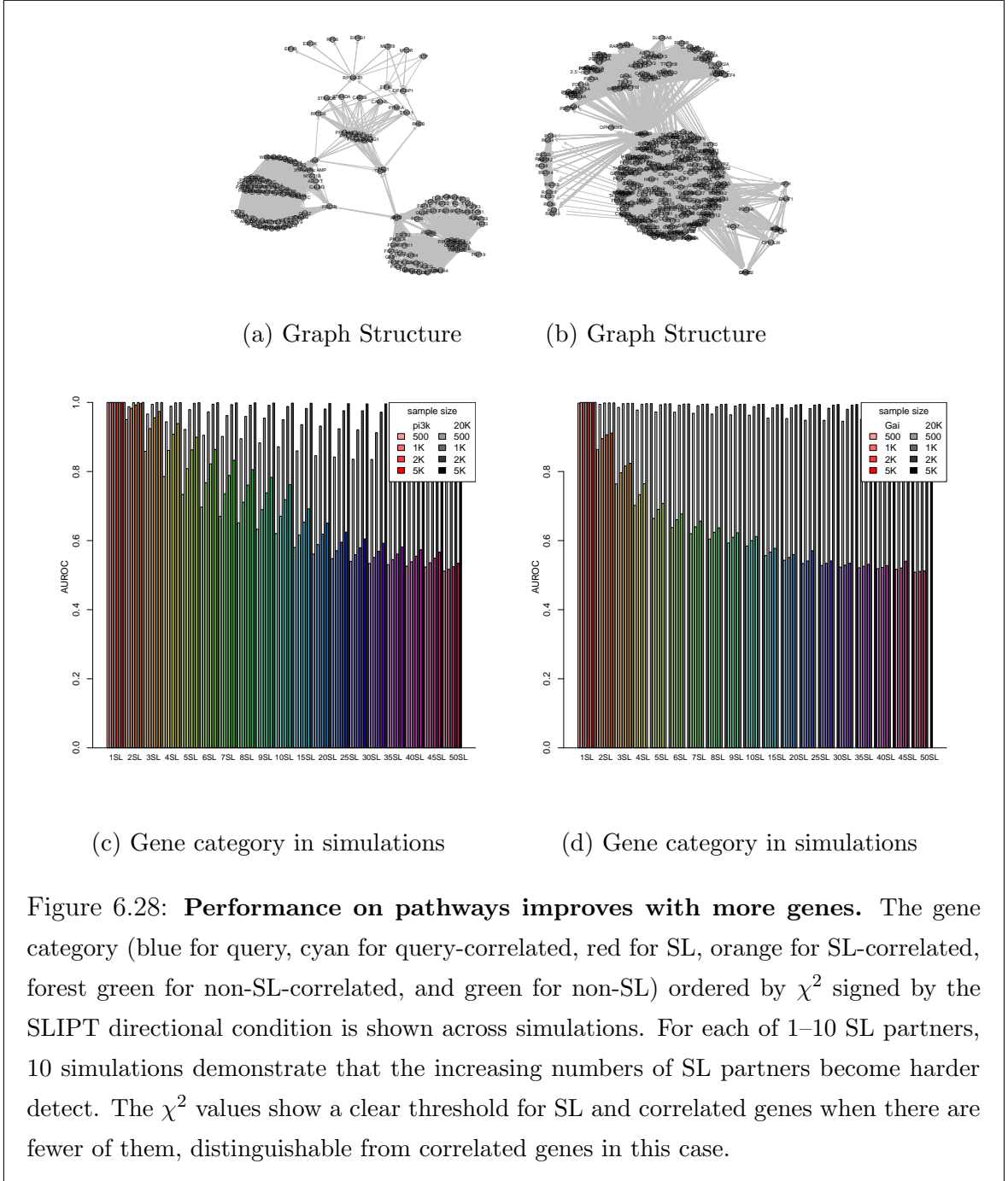
(c) Graph Structure



(d) Statistical performance

Figure 6.27: Performance of simulations including the PI3K cascade. 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.

6.3.1 Pathway Structures in a Genome



6.4 Discussion

Text

6.5 Summary

Text

Aims

- A Model of Synthetic Lethal Genes in Gene Expression Data
- Comparison of SLIPT to Alternative Approaches
- Simulations of Known Synthetic Lethal Genes within Pathway Networks

Summary

- We have designed a straight-forward rational query-based synthetic lethal detection method with the example of application to *CDH1* in cancer gene expression
- I have developed a simulation pipeline to generate continuous gene expression with pathway structure including a procedure to simulate synthetic lethality
- The simulation procedure shows that SLIPT 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 weel 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 Paediatrica*, **96**(5): 644–647.
- American Cancer Society (2017) Genetics and cancer. <https://www.cancer.org/cancer/cancer-causes/genetics.html>. Accessed: 22/03/2017.

- 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. 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., Kossenkova, 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 polyadprribose 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.
- Cancer Cell Line Encyclopedia (CCLE) (2014) Broad-Novartis Cancer Cell Line Encyclopedia. <http://www.broadinstitute.org/ccle>. Accessed: 07/11/2014.
- 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 μ -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.
- 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 knowledgebase. *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., Fedyszyn, 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.
- 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.

- 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., Scoular, R., Miller, A., and Reeve, A.E. (1998) E-cadherin germline mutations in familial gastric cancer. *Nature*, **392**(6674): 402–5.
- Guilford, P., Humar, B., and Blair, V. (2010) Hereditary diffuse gastric cancer: translation of *CDH1* germline mutations into clinical practice. *Gastric Cancer*, **13**(1): 1–10.
- Guilford, P.J., Hopkins, J.B., Grady, W.M., Markowitz, S.D., Willis, J., Lynch, H., Rajput, A., Wiesner, G.L., Lindor, N.M., Burgart, L.J., *et al.* (1999) E-cadherin germline mutations define an inherited cancer syndrome dominated by diffuse gastric cancer. *Hum Mutat*, **14**(3): 249–55.
- Guo, J., Liu, H., and Zheng, J. (2016) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res*, **44**(D1): D1011–1017.
- Hajian-Tilaki, K. (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, **4**(2): 627–635.

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009) The weka data mining software: an update. *SIGKDD Explor Newsl*, **11**(1): 10–18.
- Hammerman, P.S., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., *et al.* (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**(7417): 519–525.
- 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 gridTM (caBIGTM). *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. problèmes combinatoires at theorie des graphes. *Proc Coil 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) Identifi-

- cation 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.
- International 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, Ikonf, 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) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med*, **374**(2): 135–145.
- Lokody, I. (2014) Computational modelling: A computational crystal ball. *Nature Reviews Cancer*, **14**(10): 649–649.
- Lord, C.J., Tutt, A.N., and Ashworth, A. (2015) Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annu Rev Med*, **66**: 455–470.

- Lu, X., Kensche, P.R., Huynen, M.A., and Notebaart, R.A. (2013) Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nat Commun*, **4**: 2124.
- Lu, X., Megchelenbrink, W., Notebaart, R.A., and Huynen, M.A. (2015) Predicting human genetic interactions from cancer genome evolution. *PLoS One*, **10**(5): e0125795.
- Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., Wolf, M.K., Butler, J.S., Hinshaw, J.C., Garnier, P., Prestwich, G.D., Leonardson, A., *et al.* (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**(1): 121–137.
- Luo, J., Solimini, N.L., and Elledge, S.J. (2009) Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell*, **136**(5): 823–837.
- Machado, J., Olivera, C., Carvalh, R., Soares, P., Berx, G., Caldas, C., Sercuca, R., Carneiro, F., and Sorbrinho-Simoes, M. (2001) E-cadherin gene (*CDH1*) promoter methylation as the second hit in sporadic diffuse gastric carcinoma. *Oncogene*, **20**: 1525–1528.
- 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., Mastrogianakis, G.M., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216): 1061–1068.
- Miles, D.W. (2001) Update on HER-2 as a target for cancer therapy: herceptin in the clinical setting. *Breast Cancer Res*, **3**(6): 380–384.
- 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.
- Prahalad, 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 Matala, 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 International 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(adp-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and advanced breast cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 235–44.
- 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.