Contents

List of Figures

List of Tables

Glossary

allele	A gene variant with a specific sequence and phenotype.
bioinformatics	Statistical or computational approaches to biological data or research tools.
centrality	A network metric which identifies important vertices.
E-cadherin	Epithelial cadherin (calcium-dependent adhesion), a cell-adhesion protein encoded by *CDH1*.
edge or link	A relationship connecting a pair of elements of a graph structure or network, may be weighted or directional.
essential	A gene which is required to be functional or expressed for a cell or organism to be viable, grow or develop.
gene expression	A measure of the relative expression of each gene from the mRNA extracted from (pooled) cells.
genome	All of the DNA sequence in the genome.
genomic	The use of data from all genes in the genome.
graph or network	A mathematical structure modelling or depicting the relationships between elements.
information centrality	A network centrality metric which uses the impact of removing a vertex or node on connections in the network.
metagene	A consistent signal of expression for a collection of genes such as a biological pathway, derived from singular value decomposition.

methylation	A measure of the epigenetic regulation of DNA at CpG dinucleotide (CpG) sites.
microarray	A high-throughput technique to measure presence or abundance of nucleic acid sequences from binding to probes.
mutant	A variant or dysfunctional phenotype arising from a mutation in a gene.
mutation	A change in DNA sequence that disrupts gene function.
PageRank centrality	A network centrality metric which uses eigenvectors with a scaling factor (Brin and Page, 1998).
pathway	A series of biomolecules that produces a particular product or biological function.
RNA-Seq	The generation of transcriptome data from sequencing RNA.
shortest path	A path with the fewest possible edges which connects two particular vertices.
small world	A property of a network which is highly connected and has a low characteristic path length, derived from the mean shortest path length across all pairs of nodes.
somatic mutation	A mutation that occurs in somatic cells, during a patient's lifespan.
synthetic lethal	Genetic interactions where inactivation of multiple genes is inviable (or deleterious) which are viable if inactivated separately.
vertex or node	An element of a graph structure or network.
wild-type	A natural phenotype of a trait or the normally functional allele which encodes it.

Acronyms

ANOVA	Analysis of Variance.
AUROC	Area Under the Receiver Operating Characteristic (curve).
Bash	Bourne Again Shell.
BioPAX	Biological Pathway Exchange.
CpG	5′–C–phosphate–G–3′.
CPM	Counts Per Million mapped reads.
CPU	Central Processing Unit.
CRAN	comprehensive R archive network.
DNA	Deoxyribonucleic Acid.
ER	Estrogen Receptor.
FDR	False Discovery Rate.
HPC	High Performance Computing.
ICGC	International Cancer Genome Consortium.
mtSLIPT	Synthetic Lethal Interaction Prediction Tool (against mutation).
NeSI	New Zealand eScience Infrastructure.
PAM50	Prediction Analysis of Microarray 50.
RNA	Ribonucleic Acid.
ROC	Reciever Operating Characteristic (curve).
RPPA	Reverse Phase Protein Arrays.
RSEM	RNA-Seq by Expectation Maximization (normalisation.

siRNA	Short Interfering RNA.
SLIPT	Synthetic Lethal Interaction Prediction Tool.
Slurm	Simple Linux Utility for Resource Management.
SOCKS	Socket Secure.
TCGA	The Cancer Genome Atlas (genomics project).

Chapter 2

Methods and Resources

In this Chapter, I will outline the various existing resources and methods that were used throughout this project. This includes public data repositories, stable and development releases of software packages (primarily using the R programming environment), and custom implementation of bioinformatic methods and statistical concepts with Shell or R scripts developed for this purpose. The methods and packages that have been developed specifically for this project will be covered in Chapter 3 with supporting data and demonstration of their use .

2.1 Bioinformatics Resources for Genomics Research

2.1.1 Public Data and Software Packages

Various bioinformatics resources, such as databases and methods, have become integral parts of genetics and genomics research. Reference genomes, genotyped variants, gene expression, and epigenetics profiles are among the most commonly used resources. Gene expression data, in particular, is widely available from microarray and RNA-Seq projects, driven by data sharing, data mining, and the wider initiatives for publicly available data for enabling the scientific community to further utilise the data generated beyond a single research group or consortium (Rung and Brazma, 2013) These datasets are a valuable resource to examine the changes in gene expression occurring in cancers and the variation between samples. The potential for integrating findings from publicly available genomic data with experimental investigations has expanded with RNA-Seq datasets, including large-scale cancer genomics projects (Zhang *et al.*, 2011). This thesis presents such an investigation, enabled by the release of these datasets and tools developed to handle them.

It is now common practice for bioinformatics researchers to release open-source code or provide software packages to enable replication of the findings or further applications of the methods (Stajich and Lapp, 2006). This is part of a wider movement in software and data analysis, including the development of Linux and the R programming environment (R Core Team, 2016). In addition to the R packages hosted on comprehensive R archive network (CRAN) (CRAN, 2017), many packages specifically developed for applications in bioinformatics are hosted on the Bioconductor repositories (Gentleman *et al.*, 2004), and numerous packages in various stages of development are hosted on GitHub (`https://github.com/`). Packages from each of these resources have been used throughout this project and are cited wherever possible. Several R packages have been developed during this thesis project and publicly released on GitHub or will be released in conjunction with a publication.

2.1.1.1 Cancer Genome Atlas Data

Molecular profile data for normal and tumour samples were downloaded from publicly available sources, using the The Cancer Genome Atlas (TCGA) (TCGA, 2017) and the International Cancer Genome Consortium (ICGC) web portals (Zhang *et al.*, 2011). These include gene expression (RNA-Seq), somatic mutations, and clinical data. The versions were downloaded on the 6th of August 2015 (Release 19) and the 2nd of May 2016 (Release 20) for breast and stomach cancer respectively via the ICGC data portal (`https://dcc.icgc.org/`).

The TCGA project (Koboldt *et al.*, 2012) used widely adopted tools: "Bowtie" for alignment (Langmead *et al.*, 2009), "mapslice" to detect splice sites (Wang *et al.*, 2010), and the RNA-Seq by Expectation Maximization (RSEM) approach to quantify reads as a measure of gene expression (Li *et al.*, 2010). These are widely acceptable tools for processing RNA-Seq data which were used to produce the raw counts of mapped reads (tier 1) and normalised expression data (tier 3) publicly downloaded from ICGC and TCGA respectively. Protein expression data generated from reverse phase protein arrays (RPPA) was normalised by TCGA to dilution curves using the `SuperCurve` R package (Ju *et al.*, 2015; Neeley *et al.*, 2009).

Raw count and RSEM normalised TCGA expression data from Illumina RNA-Seq protocols were downloaded for 1177 breast samples (113 normal, 1057 primary tumour, and 7 metastases) for 20,501 genes. TCGA breast somatic mutation data for 981 samples (976 primary tumours and 5 metastases) across 25,836 genes were downloaded. These included 969 samples (964 primary tumours and 5 metastases) with corresponding RNA-Seq expression data and 19,166 genes mapped from Ensembl

identifiers to gene symbols. Of these genes, 16,156 had corresponding gene expression information. Unless otherwise stated, the raw counts were used for further processing rather than the RSEM normalised data (provided by TCGA tier 3). Somatic mutations was reported if there were non-synonymous substitutions, frameshifts, or truncations (by premature stop codons) detected which would likely disrupt the wild-type gene function. Normalised protein expression data were downloaded (as provided by TCGA tier 3), generated from RPPA for 142 antibodies targeting 115 genes for 298 TCGA breast samples.

Raw count TCGA expression data (TCGA tier 1) from Illumina RNA-Seq was downloaded for 450 stomach samples (35 normal, 415 primary tumour) for 20,501 genes. TCGA stomach mutation data was also used for 289 samples across 25807 genes, corresponding to 19436 genes with expression data.

2.1.1.2 Reactome and Annotation Data

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

2.2 Data Handling

2.2.1 Normalisation

Apart from the Prediction Analysis of Microarray 50 (PAM50) subtyping procedure (Parker *et al.*, 2009), which required RSEM normalised data (J.S. Parker personal communication), the analysis of the RNA-Seq data presented here was based on raw read count data. After some samples were removed for consistency (based on a Euclidean distance correlation matrix as described in Section 2.2.2), raw read counts were log-scaled and the final dataset was normalised as Counts per Million mapped reads (CPM), weighted by variance modelling, using the `voom` function (Law *et al.*, 2014) in the `limma` R package (Ritchie *et al.*, 2015). This procedure adjusts the data to account for differences in read count by sequencing depth between samples and length between genes.

2.2.2 Sample Triage

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

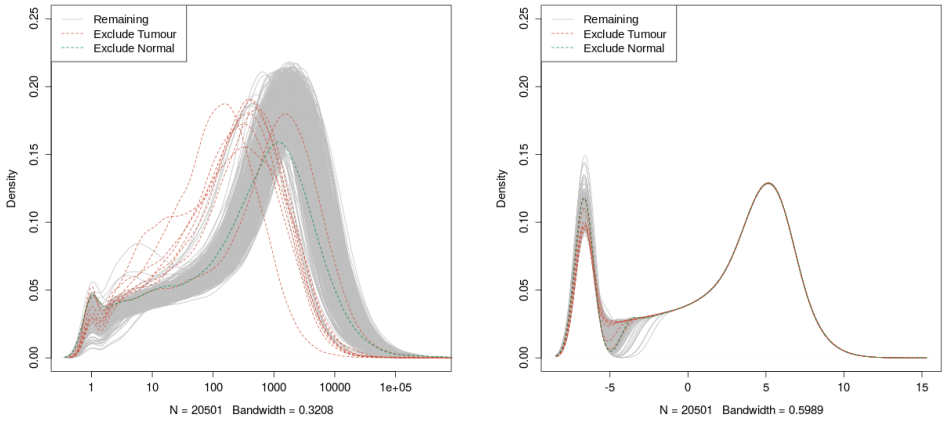
Table 2.1: Excluded samples by batch and clinical characteristics.

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

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

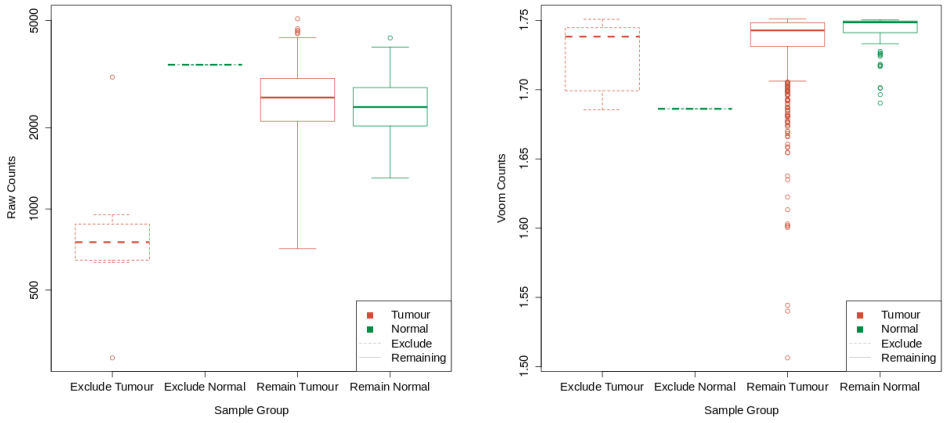
2.2.3 Metagenes and the Singular Value Decomposition

A "metagene" offers a one-dimensional summary of pathway (expression) activation or inactivation by dimension reduction of a matrix, avoiding negatively correlated genes averaging out the signal of a mean-based centroid (Huang *et al.*, 2003). Constructing pathway metagenes used gene sets for Reactome and the Gatza signatures (Gatza *et al.*, 2011, 2014) as specified above (see Section 2.1.1.2). The singular-value decomposition was performed ($X = U^T DV$, where X is the data matrix of the gene set, with genes $\times$ samples) and the leading eigenvector (first column of V) corresponding



(a) Raw counts (log-scale) (b) Voom normalised

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



(a) Mean raw counts (log-scale) (b) Mean voom normalised

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

to the largest singular value was used as a metagene for the pathway gene set. To ensure consistent directionality of metagene signals, the median of the gene set in each sample was calculated and correlated against the metagene with the (arbitrary) metagene sign adjusted as needed to conform with the majority of the gene set (i.e., positive correlation between metagene and the median-based centroid). To ensure that genes and pathways were weighted equally, metagenes were derived from a z-transformed (mean 0, standard deviation 1) dataset of gene expression and samples were scaled (by fractional ranking) for each metagene so that they were comparable on a $[0, 1]$ scale.

2.2.4 Candidate Triage and Integration with Screen Data

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

The pathway composition of the intersection was further verified by a permutation resampling analysis (as described in Section 2.3.6): the same number of genes detected by SLIPT were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1652 reactome pathways. These over-representation scores (χ^2) were compared the observed over-representation in the intersection of the SLIPT candidates, with the proportion of resamplings with higher χ^2 values used for empirical p-values of pathway composition. The χ^2 test was used as an appropriation of Fisher's exact test on a hypergeometric distribution for resampling to computationally scale pathway over-representation tests across iterations. Pathways for which no resamplings were occurred as high as the observed were reported as $p < 10^{-6}$. These empirical p-values were adjusted for multiple comparisons (False discovery rate (FDR)). Intersection size was not assumed to be constant across resamplings, so similarly with the proportion of resamplings with higher or lower intersection size were used to evaluate significance of enrichment or depletion respectively (of siRNA candidate among SLIPT candidate genes).

2.3 Techniques

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

2.3.1 Statistical Procedures and Tests

As described in sections 2.3.4 and 2.2.3, the z-transform has been used to generate z-scores in various analyses in this thesis. Each row of the dataset (x_i) is transformed into a scores (z_i) using the mean ($\bar{x}_i$) and standard deviation (s_i) of the data such that:

$$z_i = \frac{x_{ij} - \bar{x}_i}{s_i}$$

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

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

The `t.test` function (R Core Team, 2016) has also been used to implement the t-test to compare pairs of data. Where relevant, an analysis of variance (ANOVA) has been performed to report significance of multivariate predictors of outcomes, or least squares linear regression performed for the adjusted coefficient of determination (R^2) and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violin plot visualisation (Adler, 2005), rendered in R (R Core Team, 2016).

Multiple comparisons were accounted for with the Benjamini-Hochberg procedure to control the FDR unless otherwise specified (Benjamini and Hochberg, 1995). This procedure adjusts p-values to achieve an average of the proportion of false-positives among significant tests below a threshold, α. The more stringent Holm-Bonferroni (Holm) procedure (Holm, 1979) was also applied in some cases to adjust for multiple

comparisons and control the family-wise error rate which adjusts p-values so that the probability that any one of the tests is a false-positive (type-1 error) below a threshold, α.

2.3.2 Gene Set Over-representation Analysis

Gene set enrichment over-representation analysis was performed to test whether there was an enrichment of a gene set (e.g., a biological pathway) among a group of input genes. Such input genes may be predicted synthetic lethal candidates or a subset defined by clustering (in Section 2.3.3) or comparison with experimental candidates (see Section 2.2.4). Initially, these tests were performed using the GeneSetDB web tool (Araki *et al.*, 2012) hosted by the University of Auckland on the Reactome pathways (Croft *et al.*, 2014). Since the GeneSetDB tool used an older version of Reactome (version 40), it was difficult to directly compare with the results of other analysis (see sections 2.2.4 and 2.3.6) performed on version 52 (as described in Section 2.1.1.2). Thus an implementation of the hypergeometric test in R (R Core Team, 2016) was used to test for over-representation against Reactome (version 52) pathways. Pathways containing less than 10 genes or more than 500 (as performed in GeneSetDB by Araki *et al.*, 2012) were excluded before adjusting for multiple comparisons.

2.3.3 Clustering

The clustering analysis used unsupervised hierarchical clustering with complete linkage (distance calculated from the furthest possible pairing). For correlation matrices or multivariate normal parameters (e.g., Σ), the distance metric used was Euclidean distance. For empirical or simulated gene and pathway expression data correlation distance was used, calculated by $distance = 1 - cor(t(x))$ where cor is Pearson correlation and $t(x)$ is the transpose of the expression matrix.

2.3.4 Heatmap

Standardised z-scores of the data were used to plot heatmaps on an appropriate scale. Raw (log-scale) read counts or voom normalised counts per gene (as specified) were plotted as normalised z-scores on a $[-3, +3]$ blue-red scale. Similarly, correlations were plotted on a $[-1, +1]$ blue-red scale. Heatmap dendrograms were generated using the linkage method and distance specified for the clustering performed in Section 2.3.3. The `gplots` R package (Warnes *et al.*, 2015) was used to generate many of the heatmaps throughout this thesis, along with a customised heatmap function (released as `heatmap.2x`, detailed in Table 2.6 and Section 3.5.2). Where clearly specified, data

have been split into subsets with clustering performed separately on each subset with these plotted alongside each other.

2.3.5 Modelling and Simulations

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

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

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for χ^2, significance and directional criteria met for SLIPT (see Section 3.1), and an arbitrary threshold: < -0.2 and $> +0.2$ for negative correlation and correlation respectively. Sensitivity (or "true positive rate") was measured as the proportion of known synthetic lethal partners predicted to be synthetic lethal. Specificity (or "true negative rate") was measured as the proportion of known non-synthetic lethal partners predicted not to be synthetic lethal. The "false positive rate" was measured here as the proportion of known non-synthetic lethal partners out of all putative partners predicted by the detection procedure. Statistical "accuracy" is the proportion of true predictions for a detection procedure, which is both the correctly predicted known synthetic lethal partners and correctly negative known non-synthetic lethal partners.

2.3.5.1 Receiver Operating Characteristic Curves

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

2.3.6 Resampling Analysis

Resampling analyses (e.g., "permutation" analysis) are used to statistically test the significance of an observation without assuming the underlying distribution of expected test statistics (Collingridge, 2013). Instead these are derived from randomly shuffling test statistics or randomly sampling predicted candidates. For the purposes of this thesis, this involved randomly sampling genes from those tested to be analysed as putative synthetic lethal candidates. This was performed both for testing the significance of pathway composition in the intersection with experimental gene candidates (Section 2.2.4) and for assessing the significance of pathway structure among synthetic lethal candidates (Section 3.4.1.1).

These were analysed to compare the observed synthetic lethal genes against values derived from randomly sampling the same number of genes as were observed to be synthetic lethal from among the genes tested. Sampling iteratively across many resampling procedures, these resampling-based values form a null distribution that could

be expected if the null hypothesis were true. Thus the proportion of resampling-based values across these iterations that are greater than or equal to that observed, forms an empirically derived p-value to test significance.

Resampling was performed for comparison (in Section 2.2.4) with fixed experimental screen candidates (Telford *et al.*, 2015) both resampling the number of genes overlapping with the screen candidates and test statistics for pathway enrichment. Resampling analysis was also applied to shortest paths and network metrics (in Section 3.4.1.1) to test significance of directional relationships between synthetic lethal candidate genes within pathway structures.

The number of iterations determines the accuracy of these p-values. For pathway composition (in Section 2.2.4), a million iterations were performed using high performance computing (as detailed in Section 2.5.3) to provide sufficient accuracy after adjusting for multiple comparisons across pathways. For the purposes of network analysis (in Section 3.4.1.1), a thousand iterations were sufficient to reject the null hypothesis for the majority of pathways tested before adjusting for multiple comparisons, and thus further iterations were not performed.

2.4 Pathway Structure Methods

2.4.1 Network and Graph Analysis

Networks are important in considering the structure of relationships in molecular biology, including gene regulation, kinase cellular signalling, and metabolic pathways (Barabási and Oltvai, 2004). Network theory is an interdisciplinary field which combines the approaches of computer science with the metrics and fundamental principles of graph theory, an area of pure mathematics dealing with relationships between sets of discrete elements. The vast amounts of molecular and cellular data from high-throughput technologies have enabled the application of network-based and genome-wide bioinformatics analysis to examine the complexity of a cell at the molecular level and understand aberrations in cancer. This thesis uses various metrics and analysis procedures developed in Graph and Network theory to analyse graph structure of biological pathways. Where feasible, these have been implemented using the `igraph` R package with such procedures described below (Csardi and Nepusz, 2006). Custom R functions were used to perform more complex analysis and visualisation of iGraph data (as described in Section 3.5.3).

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

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

2.4.2 Sourcing Graph Structure Data

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

2.4.3 Constructing Pathway Subgraphs

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

2.4.4 Network Analysis Metrics

The existing network analysis measures applied in this thesis (as described below) used an implementation in the `igraph` R package to compute vertex degree, shortest paths, and centrality (Csardi and Nepusz, 2006). Additionally, custom features were developed for analysis of iGraph objects in R and released as `igraph.extensions` (as described in Section 3.5.3).

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

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

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

Another network centrality measure that has been previously applied to biological protein interaction networks (Kranthi *et al.*, 2013) is the "information centrality". The information centrality of a node is the relative impact on efficiency (transmission of information via shortest paths) of the network when the node is removed. Te centrality (C) (Kranthi *et al.*, 2013) for node n in graph G is defined as:

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

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

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

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

2.5 Implementation

2.5.1 Computational Resources and Linux Utilities

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

Table 2.2: Computers used during thesis

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

Table 2.3: Linux utilities and applications used during thesis

		Viao Laptop	**Lab Machine**	**Biochem Server**	**NeSI Pan Cluster**
	OS	Elementary OS Freya 0.3.2	Elementary OS Loki 0.4	Red Hat Enterprise Maipo 7.2	Cent OS Final 6.4
	Linux Kernel	3.19.0-65-generic	4.4.0-36-generic	3.10.0-327.36.2.el7.x86_64	2.6.32-504.16.2.el6.x86_64
Scripting	Shell Bash	4.3.11(1)	4.3.46(1)	4.2.46(1)	4.2.1(1)
	Shell zsh	5.0.2	5.1.1	5.0.2	5.2
Programming	Python	2.7.6	2.7.12	2.7.5	
	Java	1.8.0_101	9-ea	1.8.0_101	
	C++	4.8.4	5.4.0	4.8.5	4.4.7
Text Editor	nano	2.2.6	2.5.3	2.3.1	2.0.9
	kile (LaTeX)	2.1.3	2.1.3		
Version Control	git	1.9.1	2.11.0	1.7.1	1.8.3.1
Shell Utilities	sed	4.4.2	4.4.2	4.4.2	4.4.1
	grep	2.16-1	2.25-1	2.20	2.6.3
	nohup	8.21	8.25	8.22	8.4
Typesetting	TeX	3.1415926	3.14159265		
	TexLive (LaTeX)	2013	2015		
	PDFTeX	2.5-1	2.6		
	pandoc	1.12.2.1	1.16.0.2		
Remote Computing	Slurm scheduler				16.05.6
	OpenSSH	7.2p2	7.2p2	6.6.1	5.3p1
	OpenSSL	1.0.2g	1.0.2g	1.0.01e-fips	1.0.01e-fips
	rsync	3.1.0p31	3.1.1p31	3.0.9p30	
	Globus Online Transfer			3.1	3.1
	Cisco AnyConnect VPN		3.1.05170		
Image Processing	Inkscape	0.48.4	0.91		
	GIMP	2.8.10	2.8.16		
	ImageMagick	6.7.7.10-6			

2.5.2 R Language and Packages

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

Various scripts and packages in this thesis were developed or run in previous versions of RStudio and R but these run without error in the current version of R (and the older versions on remote machines). The R packages which were used throughout this thesis (as detailed in Table 2.5 with versions specified) were installed from the comprehensive R archive network (CRAN) (CRAN, 2017), Bioconductor (Gentleman *et al.*, 2004, version 3.4; BiocInstaller 1.24.0), or GitHub (`https://github.com/`). These packages were not updated when they would change the functionality of scripts or functions in packages, in particular imported data from annotation packages (used to define gene sets) have been saved as local files to continue using stable versions of these pathway data (across machines).

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

Table 2.4: R installations used during thesis

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

Table 2.5: R Packages used during thesis

Package	**Version Used**	**Built**	**Repository**
colorspace	1.3-2	3.3.1	CRAN
curl	2.3	3.3.1	CRAN
data.table	1.9.6	3.3.1	CRAN
dendextend	1.4.0	3.3.2	CRAN
DBI	0.5-1	3.3.1	CRAN
devtools	1.12.0	3.3.1	CRAN
dplyr	0.5.0	3.3.1	CRAN
ggplot2	2.2.1	3.3.1	CRAN
git2r	0.18.0	3.3.1	CRAN
gplots	3.0.1	3.3.1	CRAN
gtools	3.5.0	3.3.1	CRAN
igraph	1.0.1	3.3.1	CRAN
matrixcalc	1.0-3	3.3.1	CRAN
mclust	5.2.2	3.3.1	CRAN
mvtnorm	1.0-6	3.3.1	CRAN
org.Hs.eg.db	3.1.2	3.1.2	Bioconductor
openssl	0.9.6	3.3.1	CRAN
plyr	1.8.4	3.3.1	CRAN
purrr	0.2.2	3.3.1	CRAN

reactome.db	1.52.1	3.2.1	Bioconductor
RColorBrewer	1.1-2	3.3.1	CRAN
Rcpp	0.12.9	3.3.1	CRAN
ROCR	1.0-7	3.3.1	CRAN
roxygen2	6.0.1	3.3.2	CRAN
shiny	1.0.0	3.3.1	CRAN
snow	0.4-2	3.3.1	CRAN
testthat	1.0.2	3.3.2	CRAN
tidyr	0.6.1	3.3.2	CRAN
tidyverse	1.1.1	3.3.2	GitHub (hadley)
sm	2.2-5.4	3.3.1	CRAN
Unicode	9.0.0-1	3.3.2	CRAN
vioplot	0.2	3.3.1	CRAN
viridis	0.3.4	3.3.2	CRAN
xml2	1.1.1	3.3.2	CRAN
xtable	1.8-2	3.3.1	CRAN
zoo	1.7-14	3.3.1	CRAN
graphics	3.3.2	3.3.2	base
grDevices	3.3.2	3.3.2	base
cluster	2.0.5	3.3.1	base
graphics	3.3.2	3.3.2	base
grDevices	3.3.2	3.3.2	base
Matrix	1.2-8	3.3.1	base
stats	3.3.2	3.3.2	base

Table 2.6: R packages developed during thesis

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

2.5.3 High Performance and Parallel Computing

Another enabling technology for bioinformatics is parallel computing, performing independent operations using separate central processing unit (CPU) cores: this "multithreading" is widely used to increase the time to compute results. Bioinformatics is particularly amenable to this since performing multiple iterations of a simulation or testing separate genes is often "embarrassingly parallel", as CPUs completely independent of each other.

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

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

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

Chapter 3

Methods Developed During Thesis

In this Chapter, I outline the rationale and development of various methods used throughout this thesis to examine synthetic lethality in gene expression data, graph structures, models and simulations. Firstly, the Synthetic Lethal Interaction Prediction Tool (SLIPT), a bioinformatics approach to triage synthetic lethal candidate genes, will be described. This is one of the main research outputs of this thesis project and is supported by comparisons to an experimental screen from a related project and evaluation of performance on simulated data. These supporting findings will be covered in further chapters but simulation data is included to support the use and design of SLIPT. This includes the construction of a statistical model of synthetic lethality in (continuous multivariate Gaussian) gene expression data, which enables testing SLIPT upon simulated data with known synthetic lethal partners. Another key component of this simulation pipeline is the generation of simulated data from a known graph structure or simulated biological pathway (as applied in Chapter 6). The development of this simulation procedure and other statistical treatment of graph and network structures will also be covered. Various R packages have been developed to support this project, including the `slipt` package to implement the SLIPT methodology. Additional R packages for handling graph structures, simulations, and custom plotting features will be described as research outputs of this thesis, methods applied throughout, and contributions of open-source software.

3.1 A Synthetic Lethal Detection Methodology

The SLIPT methodology identifies gene expression patterns consistent with synthetic lethal interactions, between a query gene and a panel of candidate interacting partners. Gene expression is scored "low", "medium", or "high", sorting samples by tertiles (1/3-

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

The synthetic lethal prediction procedure has also been performed with somatic mutation data for the query gene. This is intended for a query gene known which is recurrently mutated, with the majority of mutations disrupting gene function (e.g., null or frameshift mutations). A synthetic lethal interaction is predicted (as shown in Figure 3.2) when (i) the χ^2 test is significant; (ii) observed mutant-query, low-candidate samples are less frequent than expected; and (iii) observed mutant-query,

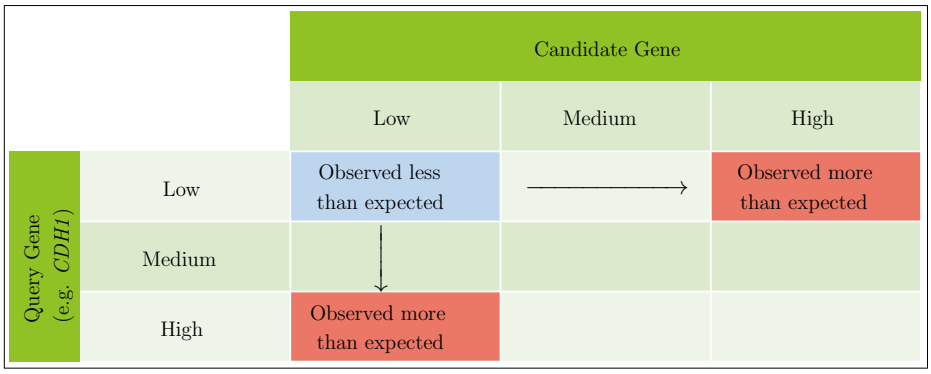


Figure 3.1: **Framework for synthetic lethal prediction.** SLIPT was designed to identify candidate interacting genes from gene expression data using the χ^2 test against a query gene. Samples are sorted into low, medium, and high expression quantiles for each gene to test for a directional shift. A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

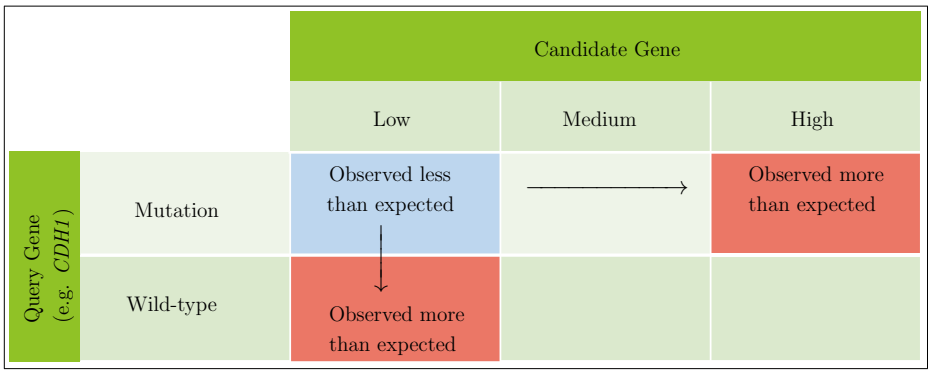


Figure 3.2: **Synthetic lethal prediction adapted for mutation.** SLIPT was also adapted to identify candidate interacting genes using (somatic) mutation data of the query gene in the χ^2 test. Samples are sorted into low, medium, and high expression quantiles for each candidate gene and tested for a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

high-candidate and wild-type-query, low-candidate samples are more frequent than expected.

The SLIPT methodology can be performed on expression data, including pathway metagenes (as generated in Section 2.2.3). The application of the SLIPT methodology on public gene expression data will be supported with simulation results (in Section 3.3 and Chapter 6), including comparison to other statistical methods. SLIPT results for *CDH1* were compared experimental screen results in a breast cell line (Telford *et al.*, 2015). Primary screen results are discussed in Section 4.2 and secondary screen results are presented in Section 4.2.4.

3.2 Synthetic Lethal Simulation and Modelling

A statistical model of synthetic lethality was developed to generate simulated data and to evaluate the SLIPT procedure. This section describes the synthetic lethal model and the simulation procedure for generating gene expression data with known synthetic lethal partners. Simulation results, to support usage of the SLIPT methodology throughout this thesis, will be presented in Section 3.3. The simulation procedure will

also be applied in Chapter 6, including in combination with simulations from graph structures (as described in Section 3.4.2).

3.2.1 A Model of Synthetic Lethality in Expression Data

A conceptual model of synthetic lethality was devised (see Figure 3.3), which will be used to build a statistical model of synthetic lethal gene expression and to simulate expression data for assessing various potential synthetic lethal prediction methods, including SLIPT. In the model, synthetic lethality occurs between genes with related functions, as a cell death phenotype, when these functions are inactive.

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

The first condition of the synthetic lethal model is that if there are only two synthetic lethal genes (e.g., *CDH1* and one SL partner), then they will not both be non-functional in the same sample (in an ideal model). Gene function is thus determined for each sample in a model of synthetic lethality with the proportion of samples which are functional or non-functional for a gene being arbitrary. Whether a gene is functional can similarly be modelled by an arbitrary threshold of continuous and normally distributed gene expression data to define gene function (as shown in Figure 3.4). For the purposes of modelling synthetic lethality in cancer expression data, a threshold of the 30^{th} percentile of the expression levels was used because approximately 30% of samples analysed had *CDH1* inactivation (mutations) in breast cancer (Koboldt *et al.*, 2012). This was generalised for a model of the proportion of samples inactivated for each gene. The threshold of the 0.3 quantile was used in simulations dervied from this

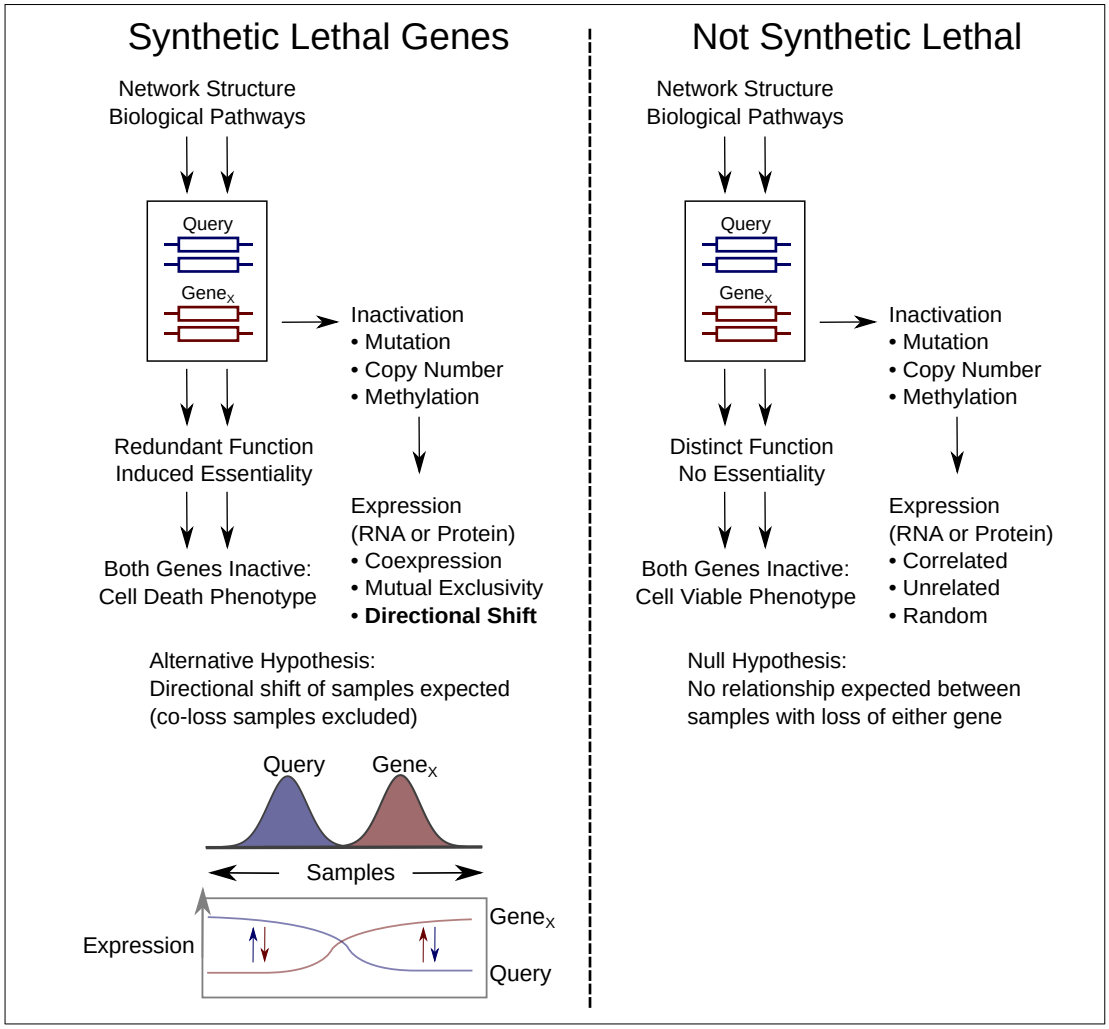


Figure 3.3: **A model of synthetic lethal gene expresion.** A conceptual model of synthetic lethal interactions between a Query gene and partner gene (G_X). Genes that are synthetic lethal may not both be non-functional in the same sample without another gene compensating for the loss of function. This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

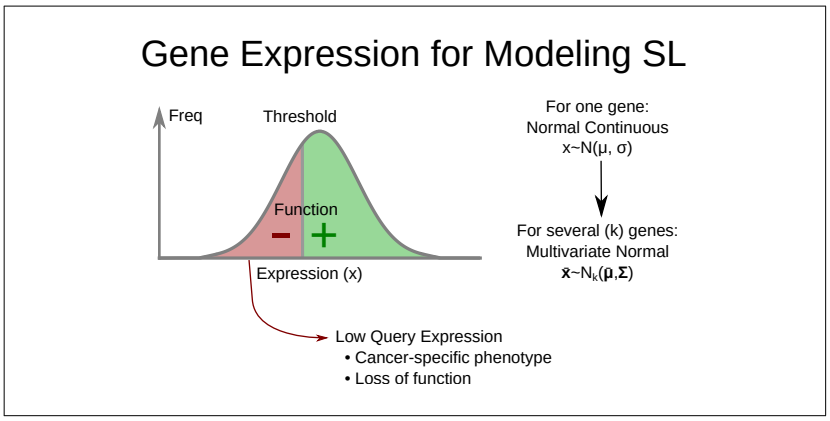


Figure 3.4: **Modelling synthetic lethal gene expression.** When modelling synthetic lethal interactions between a Query gene and partner genes (G_X and G_Y) above, cellular viability requires that at least of genes is not inactivated. As a model of loss of function, genes are regarded as non-functional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

model throughout this thesis. In this ideal case, no samples lowly expressing both of these genes are expected to be observed. While this was not the case, that is to be expected as it is unlikely that only two genes will have an exclusive synthetic lethal partnership.

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

The signal (in terms of gene expression data) will be weaker for this latter case and this model has the more stringent assumption that all synthetic lethal partner genes interact with each other: that only one of these must be expressed to satisfy

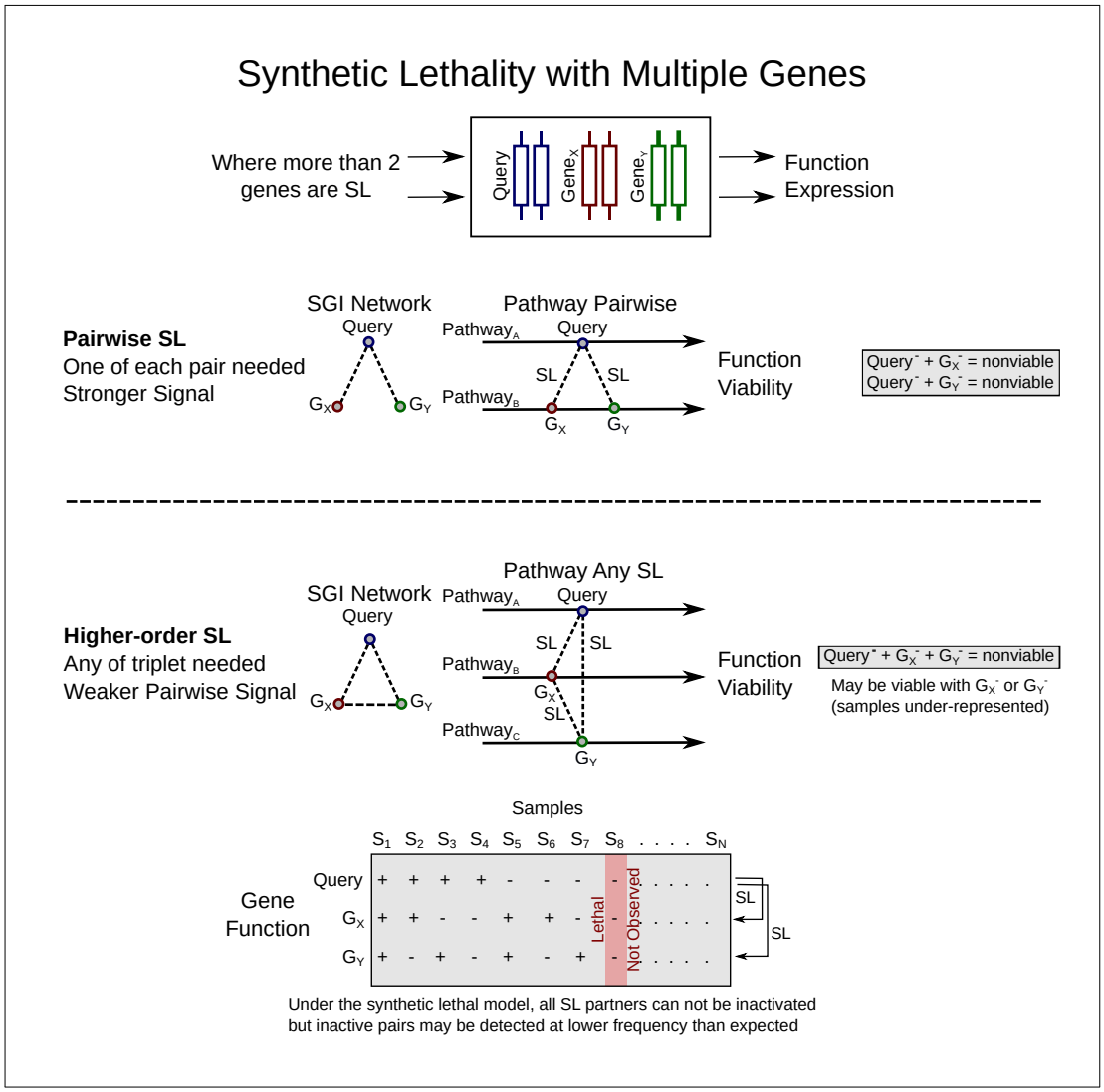


Figure 3.5: **Synthetic lethality with multiple genes.** Higher order synthetic lethal interactions may occur between 3 or more genes, affecting the simulated expression (or synthetic lethal predictions) even if undetected when observed pairwise. Consider interactions between a Query gene and two partner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

the model of synthetic lethality. In this model, any of the synthetic lethal genes in a higher-order interaction are able to perform the essential function of the others, allowing for higher-level synthetic lethal partners to compensate for loss a synthetic lethal gene pair. While samples that express low levels of the synthetic lethal gene pairs will be under-represented, they may not be completely absent from the dataset, due to these higher-level interactions. In the example of three synthetic lethal genes (shown in Figure 3.5), only one of the genes involved in the higher-order synthetic lethal interaction is required for cell viability. For synthetic lethal pairs, only a subset of these samples will be inviable (i.e., removed from simulated data), leading to an under-representation.

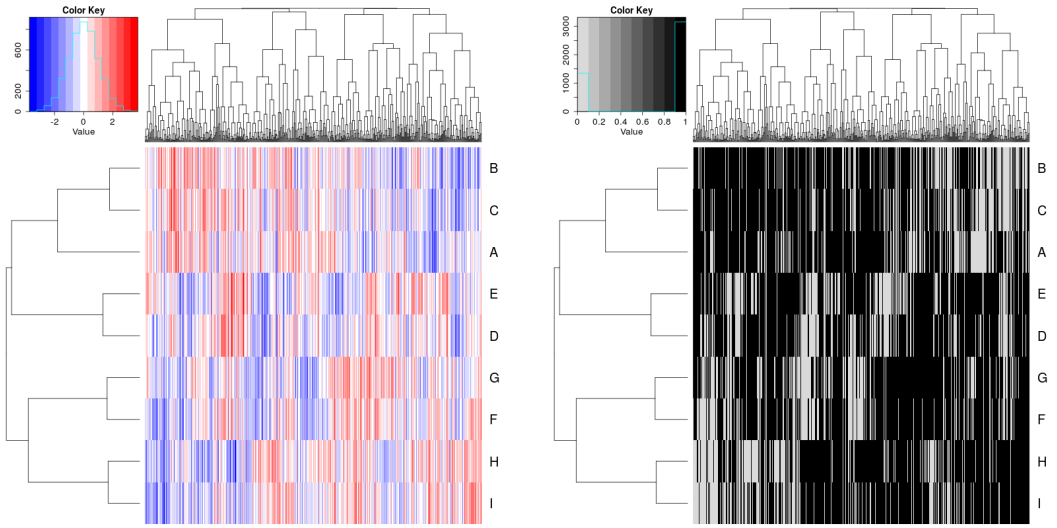
Samples were not actually removed from a simulated dataset, rather the expression and function of the query gene is generated across samples separately from the pool of potential partner genes. The query gene data was matched to simulated samples (as shown in Figure 3.7) satisfying the synthetic lethal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

3.2.2 Simulation Procedure

Simulations were developed to generate normal distributions of expression data and define gene function with a threshold cut-off. While gene function was used as an intermediary step in modelling synthetic lethal genes in expression data, the normal distribution was sampled for simulated data to represent normalised empirical gene expression data for which SLIPT (and other methods) will be applicable.

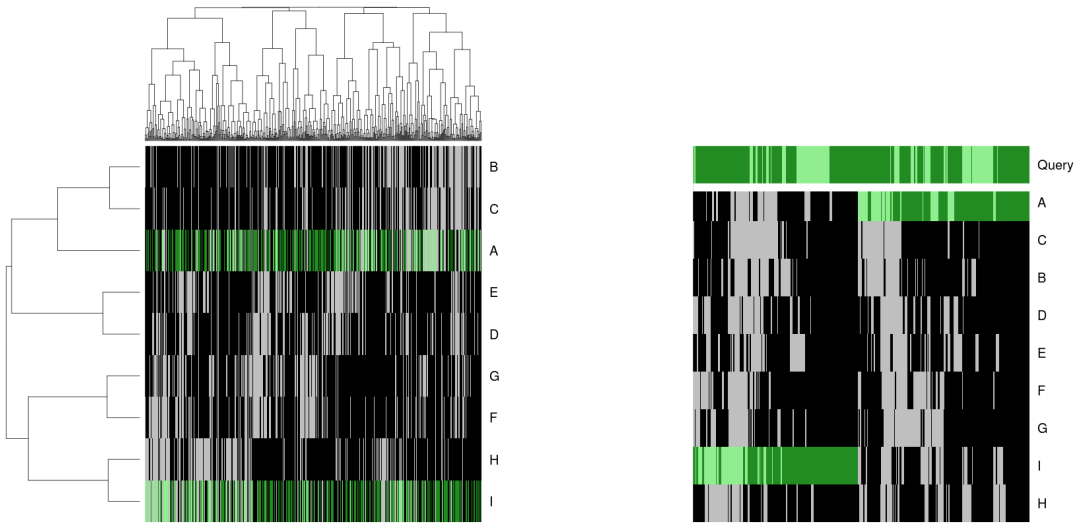
Sampling a distribution for expression profiles has the advantage of enabling simulating correlation structures with the multivariate normal distribution, using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016). The parameter Σ, the covariance matrix, defines the correlation structure between the simulated genes being sampled. With a diagonal of one, this Σ matrix simulates genes with a standard deviation of one and the covariance parameters between them are the correlations between each gene. In Figure 3.6, an example of such a simulated multivariate normal dataset is shown with the functional threshold applied.

Once a simulated dataset has been generated, the samples were compared by gene function (as derived from a functional threshold). The known underlying synthetic lethal partners were selected within the dataset and a query gene was generated by sampling from the normal distribution. These were matched (as shown for two synthetic



(a) Simulated expression matrix (b) Corresponding gene function calls

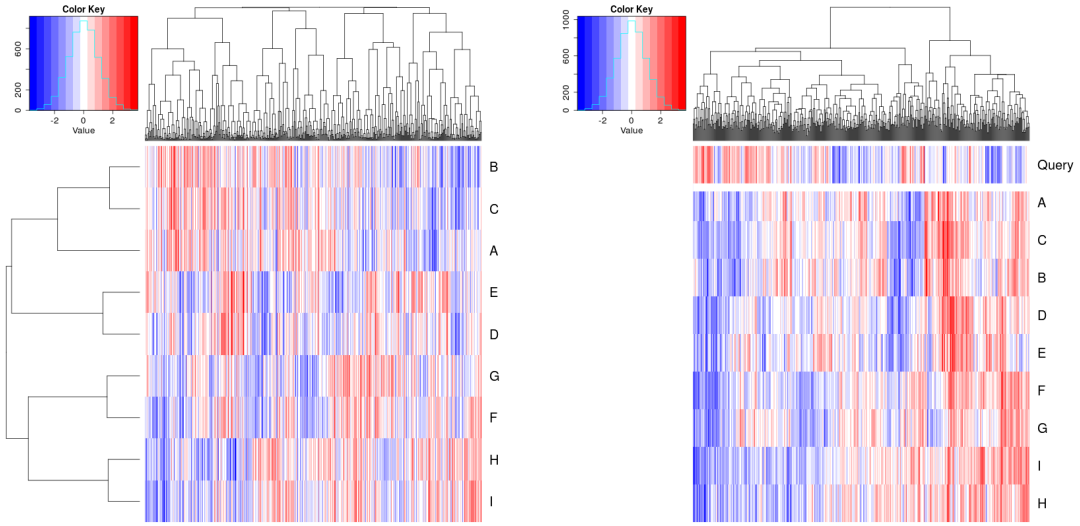
Figure 3.6: **Simulating gene function.** A simulated dataset with samples (columns) and genes A–I (rows) was transformed from a continuous (coloured blue–red) scale to a discrete matrix of gene function (black for functional levels and grey for non-functional).



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

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

lethal partners in Figure 3.7) such that the synthetic lethal condition was met: at least one of the synthetic lethal partner genes and the query gene are functional in any particular cell. The samples are ordered by functional data (without assuming correlation of underyling expression values) with the query gene in one direction and the remaining dataset ordered by the selected synthetic lethal partner.



(a) Initial expression matrix (b) Simulated synthetic lethal dataset

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

This procedure produces a simulated dataset where samples with a non-functional query gene have at least one functional partner gene. Similarly, the query gene is functional in all samples where all of the synthetic lethal partner genes are non-functional. In this procedure, a dataset has been generated with known synthetic lethal partners (see Figure 3.8) with few assumptions about the relationships between the each synthetic lethal pair (allowing compensating functions from higher-order interactions). This procedure has been designed to have the most stringent (least detectable) synthetic lethal relationships, where higher-order interactions are possible for the purposes of testing pairwise detection procedures such as SLIPT.

3.3 Detecting Simulated Synthetic Lethal Partners

The synthetic lethal detection methodology (SLIPT), as described in Section 3.1, was evaluated with simulated data containing known synthetic lethal partners, generated using the procedure described in Section 3.2.2. Simulations were performed to demonstrate the methodology and support its use throughout this thesis. These simulations were performed by sampling from statistical distributions, including the multivariate normal distribution with correlated blocks of genes, generated by Σ matrices such as those shown. A more complex multivariate normal sampling procedure based on pathway graph structures, as described in section 3.4.2, was used for further investigations in Chapter 6.

3.3.1 Binomial Simulation of Synthetic Lethality

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

$$X \sim Bin(n, p)$$

Once a query gene with synthetic lethal partners has been added, these functional levels were passed directly into SLIPT as "low" and "high" samples.

The simulation procedure was performed with 20,000 total genes (as occurs in expression datasets) with a variable number of true synthetic lethal partners and 500, 1000, 2000, or 5000 samples. Each ROC curve was derived from the results of 10,000 replicate simulations. The statistical performance (as shown in Figure 3.9) of the χ^2-derived p-value declined towards random predictions (an AUROC of 0.5) with an more underlying synthetic lethal partners to detect. However, increased sample size somewhat mitigated this decline, as expected with a statistical predictor, particularly for moderate numbers of synthetic lethal partners.

Simulations using this binomial model of synthetic lethality were simplistic but informed the development of more complex simulations including expression and correlation structures. It did not represent the data that SLIPT will be applied to but the binomial simulations demonstrated that SLIPT is able to distinguish small numbers of synthetic lethal partners in a simulated system with behaviours expected with respect to sample size. This supported further development of the synthetic lethal model

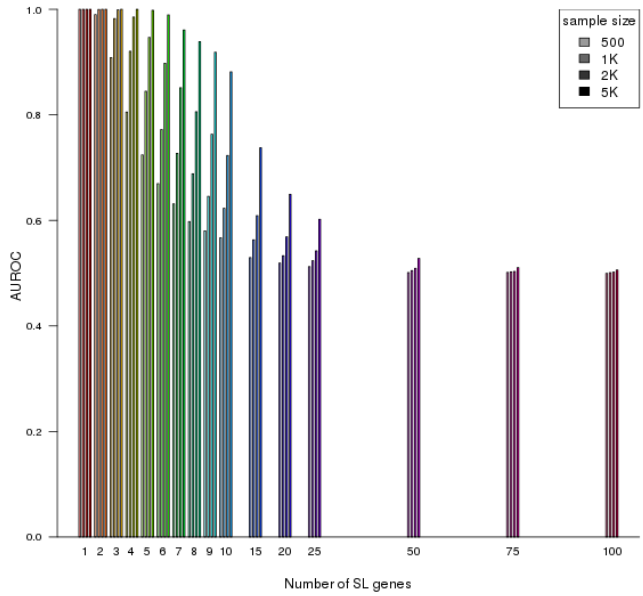


Figure 3.9: **Performance of binomial simulations.** Gene function was simulated by binomial sampling and tested for synthetic lethality by SLIPT. Statistical performance declined with additional synthetic lethal partners but this was mitigated by increased sample sizes.

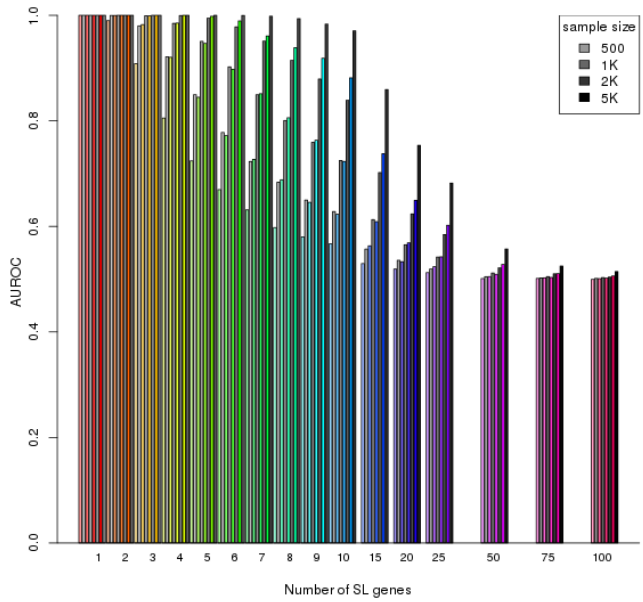


Figure 3.10: **Comparison of statistical performance.** Binomial simulation of synthetic lethality (in colour) in comparison to multivariate normal simulations (in greyscale), in which SLIPT consistently had higher performance across parameters.

and simulation pipeline (as described in Section 3.2) using the multivariate normal distribution.

The multivariate normal simulation procedure more closely recapitulates the (normalised) expression data that SLIPT was intended for and enables the methodology procedure to be tested without requiring modifications (in Section 3.3.2). Sampling continuous expression values from a normal distribution enabled the expression threshold for gene function to differ from the categorical "low" and "high" expression binning performed by SLIPT (as discussed in Section 3.2.1). The SLIPT procedure does not assume a known threshold for expression and instead uses expression as an estimate of gene function which does not compromise the statistical performance of the SLIPT in the multivariate normal simulation. The performance was an improvement over the binomial simulation procedure (shown in Figure 3.10) across simulation parameters in an equivalent simulation (without correlation structure). This multivariate normal model is also more refined since it defines the synthetic lethal condition, to ensure that at least one synthetic lethal partner was active in query-deficient samples, without disrupting the proportion of samples with each gene being functional.

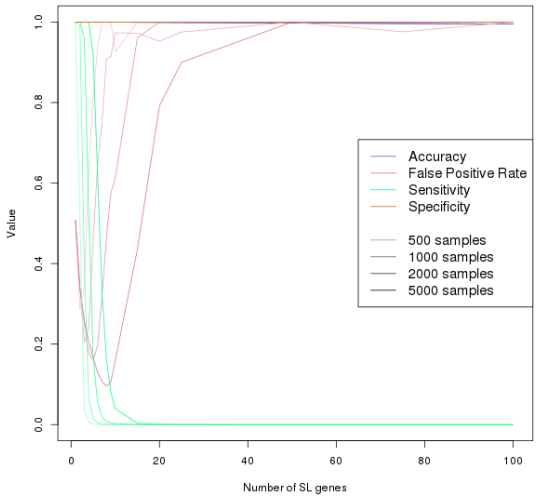
3.3.2 Multivariate Normal Simulation of Synthetic Lethality

The multivariate normal simulation procedure was initially performed using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016) (as described in Section 3.2) without correlation structure. Expression was sampled from multivariate normal distribution with a mean ($\mu = 0$), standard deviation ($\sigma = 1$), and no correlation between genes ($r = 0$):

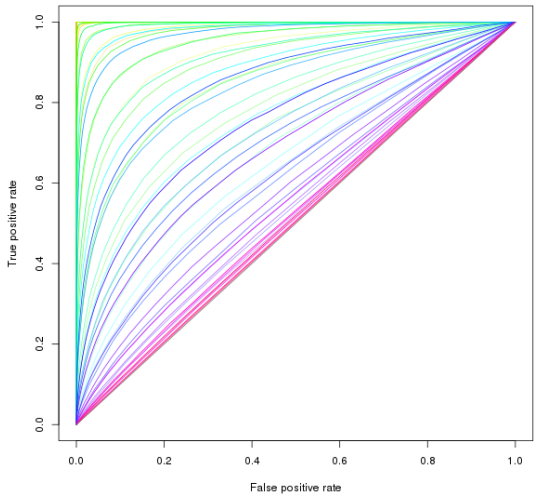
$$X \sim N(\bar{\mu}, \Sigma).$$

Once a query gene with synthetic lethal partners has been added, the simulated expression values were tested by SLIPT, as described in Section 3.1.

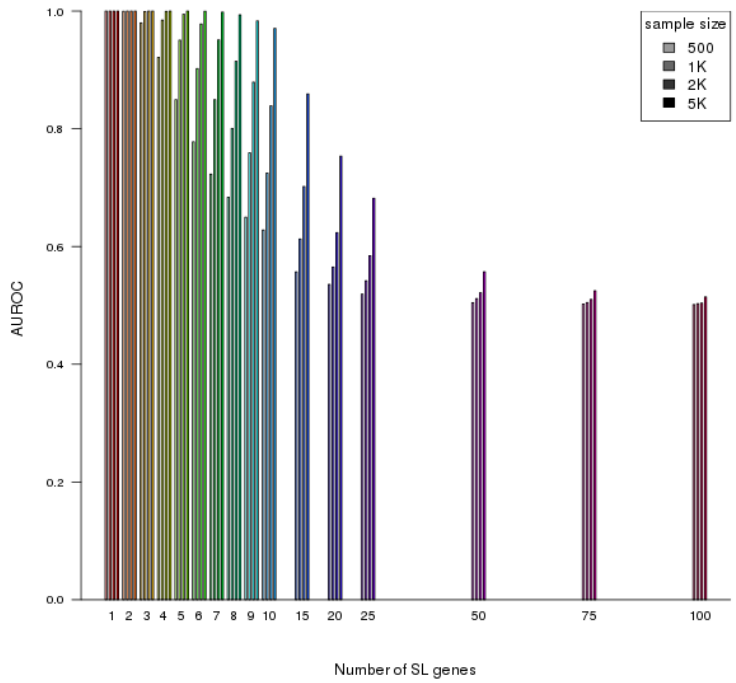
The statistical accuracy of SLIPT as a binary classifier was high across simulations of a full dataset of 20,000 genes (shown in Figure 3.11a). Using the χ^2-derived p-value as a threshold for prediction, this was largely due to high specificity: the majority of non synthetic lethal genes were distinguished from the underlying synthetic lethal genes. Thus the SLIPT methodology performed better with larger datasets with more expected negatives and the results of simulations of smaller numbers of genes (e.g., the graph structures analysed in Section 6.2.1) can be applied to larger datasets, where they are expected to perform comparably or better with a lower false negative rate



(a) Statistical evaluation



(b) ROC curves



(c) Statistical performance (AUROC)

Figure 3.11: **Performance of multivariate normal simulations.** Simulation of synthetic lethality was performed by sampling from a multivariate normal distribution (without correlation structure). Performance of SLIPT declined with increasing numbers of synthetic lethal partners but this was mitigated by increased sample sizes (in darker colours). This occurred as the sensitivity decreased with a greater number of true positives to detect, which lead to a trade-off in accuracy as seen in a trough for false positive rate and the ROC curves.

(as shown in Sections 6.2.4 and 6.3.2). Accordingly, key results will be supported by replication with larger numbers of non synthetic lethal genes added to the simulations.

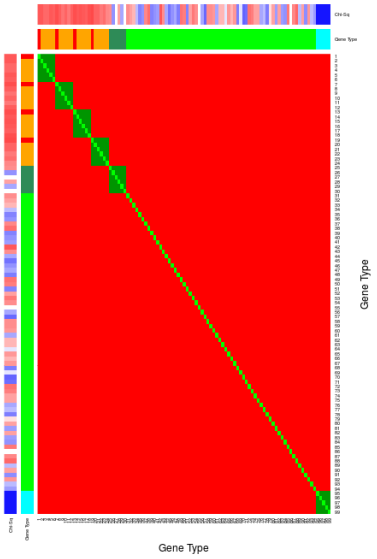
The sensitivity of SLIPT as a binary classifier of synthetic lethality (in Figure 3.11a) declined with higher numbers of synthetic lethal genes to detect, although this is somewhat mitigated by higher sample sizes. The minority of true synthetic lethal partners are more difficult to distinguish when there are more of them (with a weaker expression signal from each). While a reduction of the false positive rate could be achieved for moderate numbers of underlying synthetic lethal partners, the number of partners to be detected in analyses of expression data is unknown. However, this simulation procedure is amenable to assessing the performance of SLIPT across simulation parameters, graph structures, and comparisons to other approaches (as presented in Chapter 6).

Not all of the genes detected by SLIPT were true synthetic lethal genes but they were among the strongest candidates and SLIPT had higher performance with fewer underlying synthetic lethal genes to detect. These results support a focus on pathway analyses, in particular, the selection of pathways for further investigation. Pathway over-representation analysis was performed to detect functional groups recurrently detected by SLIPT since individually detected gene candidates were not necessarily synthetic lethal. The detection of functionally related genes (in Chapter 4) supports the role of a pathway in synthetic lethal relationships. The use of pathway metagenes can reduce the number of potential pathways, compared to genes, to help identify synthetic lethality. These approaches were both applied in Chapter 4 to identify the synthetic lethal pathways of *CDH1*. Pathways are also more likely to replicate across experimental models, as demonstrated by Dixon *et al.* (2008).

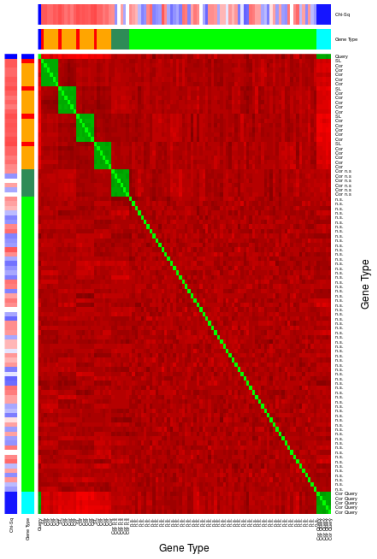
The ROC curves showed that SLIPT is subject to a near equal trade-off between sensitivity and specificity across threshold values (in Figure 3.11b). The lower sensitivity and higher specificity with a binary classification (in Figure 3.11a) results from stringent testing by SLIPT with FDR adjusted p-values. The area under these curves (AUROC) was used to compare statistical performance (in Figure 3.11c), which had lower performance for more underlying synthetic lethal partners, and higher performance for larger sample size in multivariate normal simulations.

3.3.2.1 Multivariate Normal Simulation with Correlated Genes

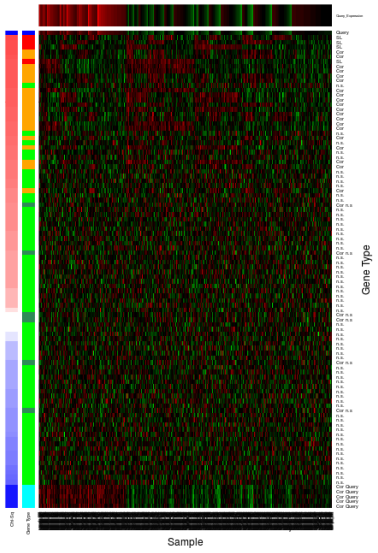
Correlation structures were added to the simulation procedure (with the Σ matrix, as discussed in Section 3.2), using correlated blocks of genes (as shown in Figure 3.12a). These correlated blocks represent genes with correlated expression, such as co-regulation or shared membership of biological pathways. The example (in Figure 3.12) shows four



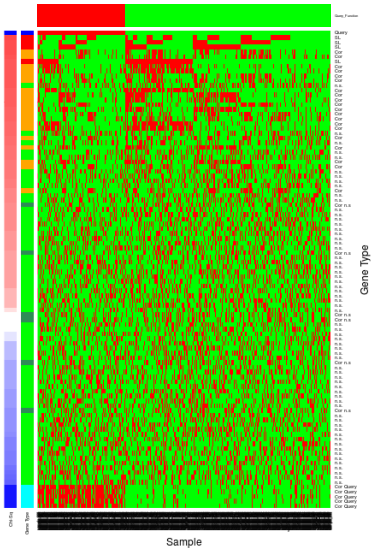
(a) Input Σ matrix parameter



(b) Simulated correlation matrix

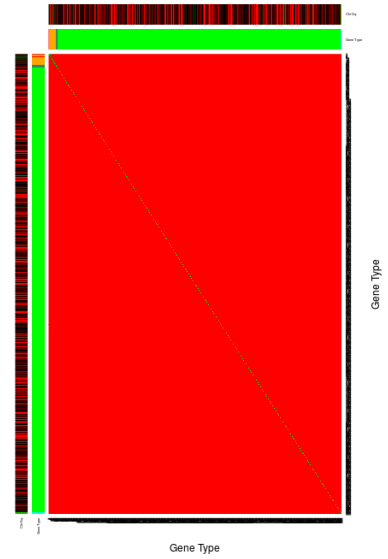


(c) Simulated gene expression

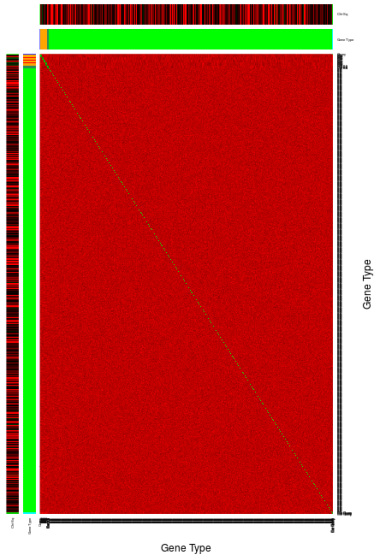


(d) Simulated gene function

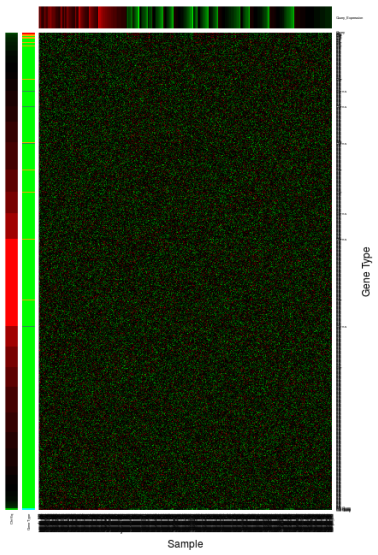
Figure 3.12: **Simulating expression with correlated gene blocks.** A Σ matrix (a) was used generate 100 genes with a multivariate normal distribution, including correlated blocks of genes ($r = 0.8$) with correlation (b) similar to Σ, on a red–to–green scale. The annotation for genes gives the χ^2 (in blue for in the direction of SLIPT or red otherwise) and the gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non synthetic lethal-correlated, and green for non synthetic lethal). The simulated gene expression (c) and function (d) generated were ordered by χ^2 showing the functional structure of synthetic lethal genes and that they were among the strongest SLIPT results.



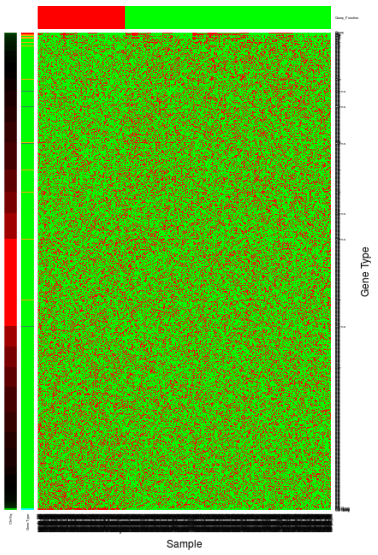
(a) Input Σ matrix parameter



(b) Simulated correlation matrix



(c) Simulated gene expression



(d) Simulated gene function

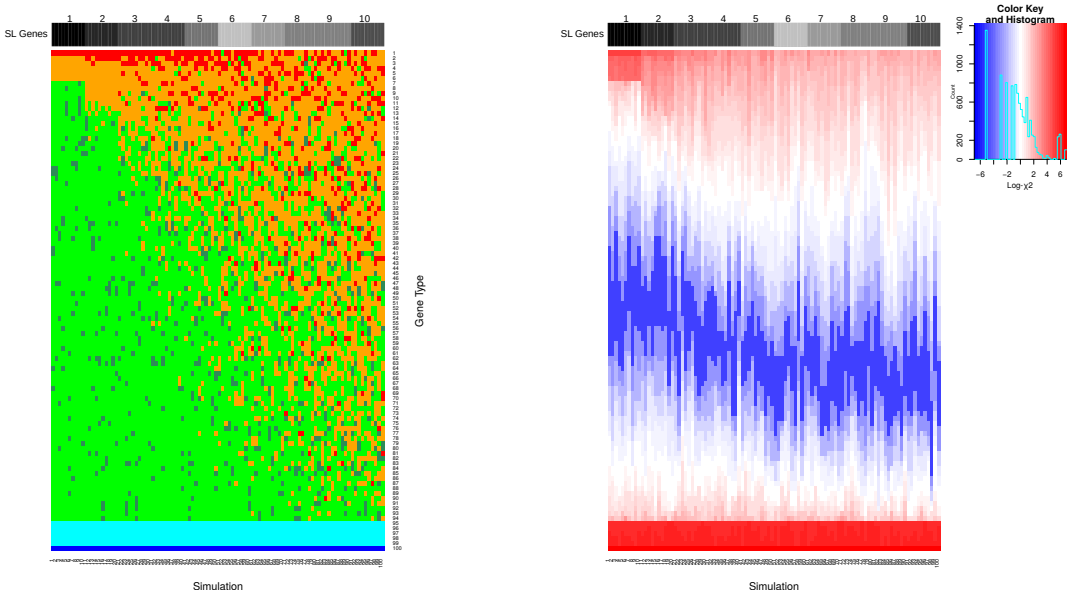
Figure 3.13: **Simulating expression with correlated gene blocks.** Using the Σ matrix (a), sampling 1000 genes from a multivariate normal distribution produced (b) correlated blocks of genes (correlated by 0.8) on a red–to–green scale. The simulated gene expression (c) and function (d) generated were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

synthetic lethal genes (out of 100), each with five correlated genes that are not themselves synthetic lethal partners of the query gene. These simulations address whether synthetic lethal genes are distinguishable from correlated partners. The Σ matrix produced a similar correlation structure (Figure 3.12b) and expression profiles (Figure 3.12c). Apart from correlated blocks of genes ($r = 0.8$), the remaining genes had small variations due to random sampling. The structure of the dataset, particularly between synthetic lethal genes and the query, was evident in the simulated gene expression (Figure 3.12c) and function (Figure 3.12d). When these genes were ordered by the SLIPT results, the synthetic lethal genes were highly ranked, a the majority of them were distinguishable from highly correlated genes.

The use of correlation structure was applied to larger datasets, such as the 1000 genes shown in Figure 3.13. Synthetic lethal genes were highly ranked by SLIPT and were often distinguishable from correlated genes. As previously discussed in Section 3.3.2, these synthetic lethal genes were still detectable among a larger number of non synthetic lethal genes, and the SLIPT methodology performed better on large datasets.

These plots (Figures 3.12 and 3.13) used similar correlated blocks with a non synthetic lethal gene (true negative) and the query gene (which is not synthetic lethal with itself). Neither of these were synthetic lethal but they could potentially affect performance methodology, particularly the specificity, as correlated non synthetic lethal genes may be distinguishable from synthetic lethal genes. The non synthetic lethal correlated block of genes had no impact on synthetic lethal detection but the query correlated genes were important (as shown in Sections 3.3.2.2 and and 6.1.1.1).

The simulations of gene expression data (with 100 genes) with correlations structure were used to examine the variation between detection in different samples and varying the number of underlying synthetic lethal partners. A small number of simulations (10 for each) are shown to demonstrate the variation between replicate simulations from iterative sampling from the same multivariate normal distribution (in Figure 3.14). These simulations showed that synthetic lethal genes were highly ranked by SLIPT when there are few of them and these were relatively consistent across replicate simulations. However, they were less consistent for higher numbers of synthetic lethal partners to detect and were more difficult to distinguish from other genes, particularly those correlated with them. Similarly, the χ^2 values showed clear thresholds for synthetic lethal and correlated genes in simple simulations but these were more gradual for higher numbers of synthetic lethal partners.



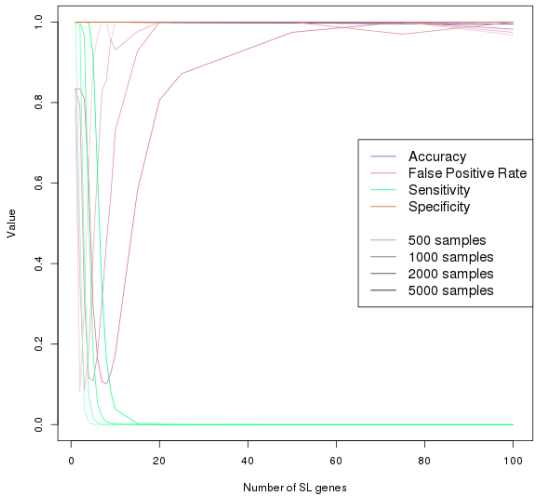
(a) Gene category in simulations

(b) Corresponding χ^2 values

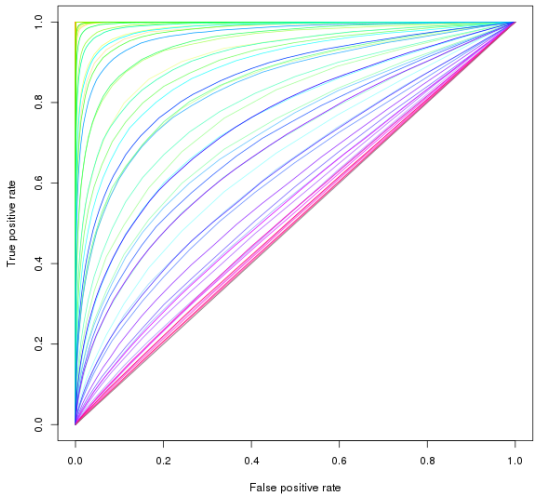
Figure 3.14: **Synthetic lethal prediction across simulations.** The gene category (a) ordered by χ^2 and the SLIPT directional condition is shown across simulations (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non synthetic lethal-correlated, and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

While the synthetic lethal genes were detected in simple simulations (in Figure 3.14), ROC analysis was performed to determine whether they were robustly detectable and to make further comparisons. These results (in Figure 3.15) were similar to simulations without correlation structure. As a binary classifier, SLIPT had low sensitivity for higher numbers of synthetic lethal partners to detect and high specificity with the vast majority of non synthetic lethal genes (for 20,000 genes). This was reflected in a similar reduction in statistical performance for higher numbers of synthetic lethal partners and a higher performance with higher sample size. Overall, the statistical performance was no different to simulations without correlation structure (as shown in Figure 3.16).

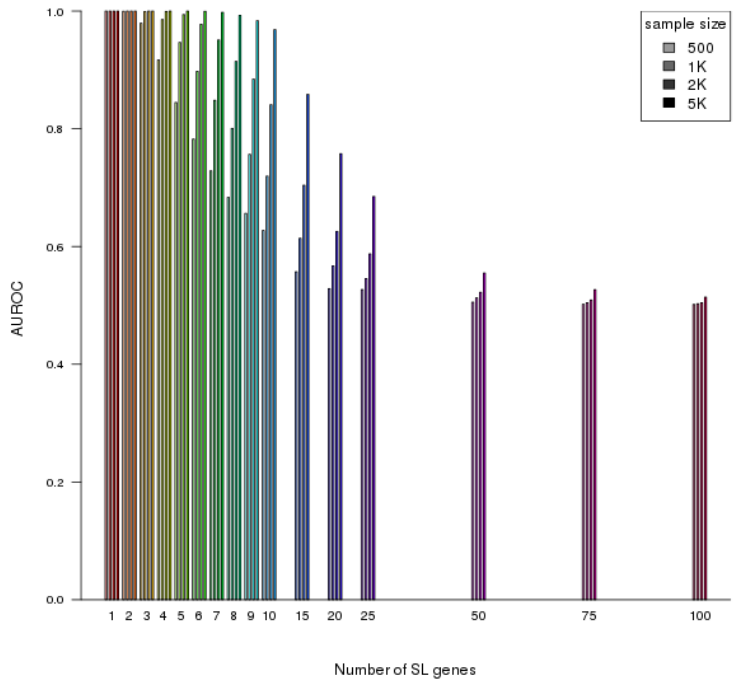
SLIPT was robust across correlation structures and is applicable to gene expression data, with pathway structures and correlations. These correlation structures were not intended to model specific biological pathways or represent them but showed potential



(a) Statistical evaluation



(b) ROC curves



(c) Statistical performance (AUROC)

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

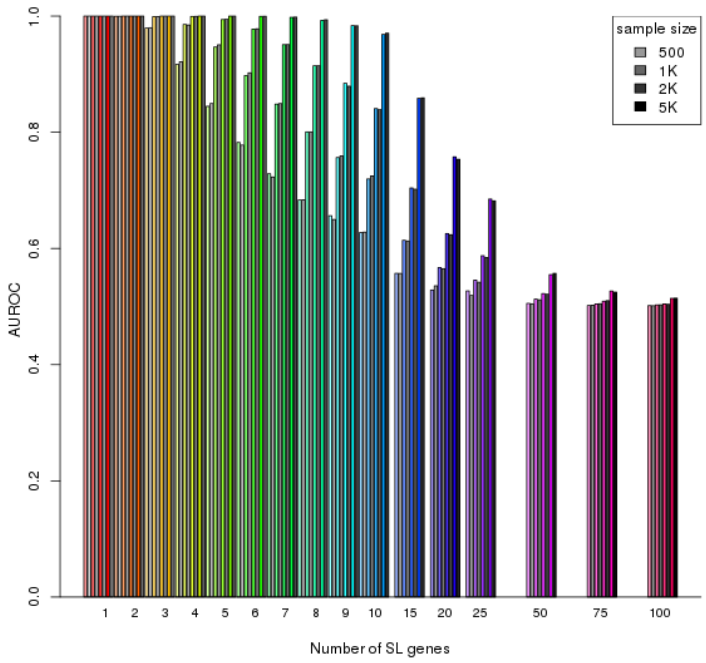


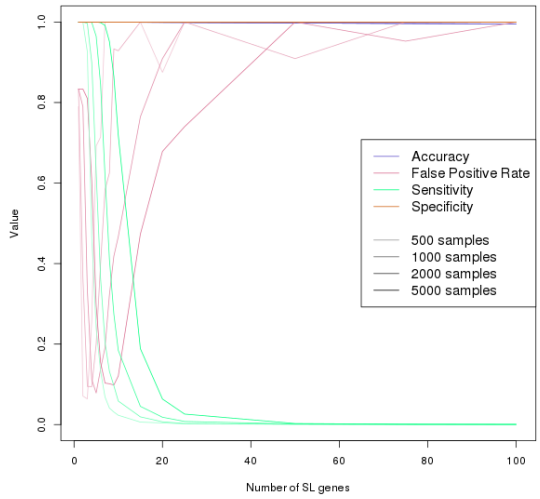
Figure 3.16: **Comparison of statistical performance with correlation structure.** Multivariate simulation of synthetic lethality with correlation structure (in colour) has comparable performance to simulation without correlations (in greyscale) with known synthetic lethal partners across parameters.

impact of correlation structure on the performance of SLIPT using highly correlated ($r = 0.8$) gene blocks. More complex correlation structures, such as genes positively correlated with the query gene and derived from pathway graph structures (as described in 3.4.2) were examined further in Sections 3.3.2.2 and 6.2.1 respectively.

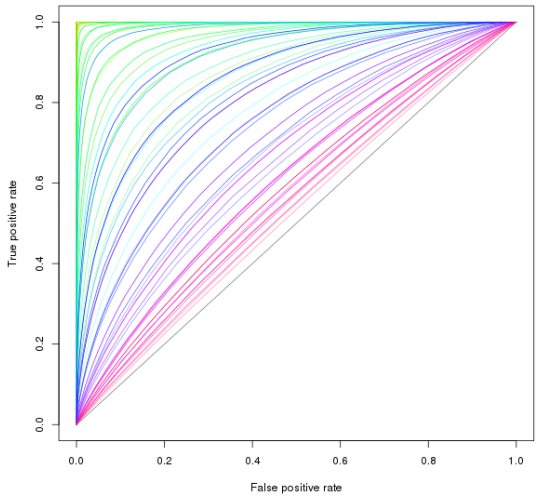
In particular, genes correlated with true synthetic lethal genes had little impact on the performance of SLIPT detection: synthetic lethal genes were as distinguishable from correlated genesas as they are from true negative genes. Genes correlated with synthetic lethal partners did not interfere with the detection of true synthetic lethal genes, although they were often ranked next below them and may support synthetic lethal pathways by having related gene functions.

3.3.2.2 Specificity with Query-Correlated Pathways

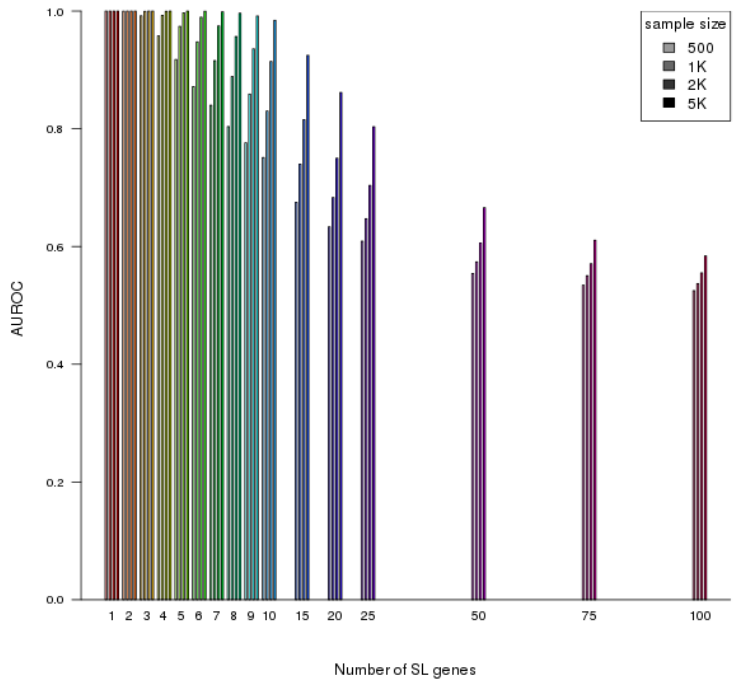
Correlation structures were also considered for non synthetic lethal genes that were (positively) correlated genes with the query gene. Specifically, five highly correlated



(a) Statistical evaluation

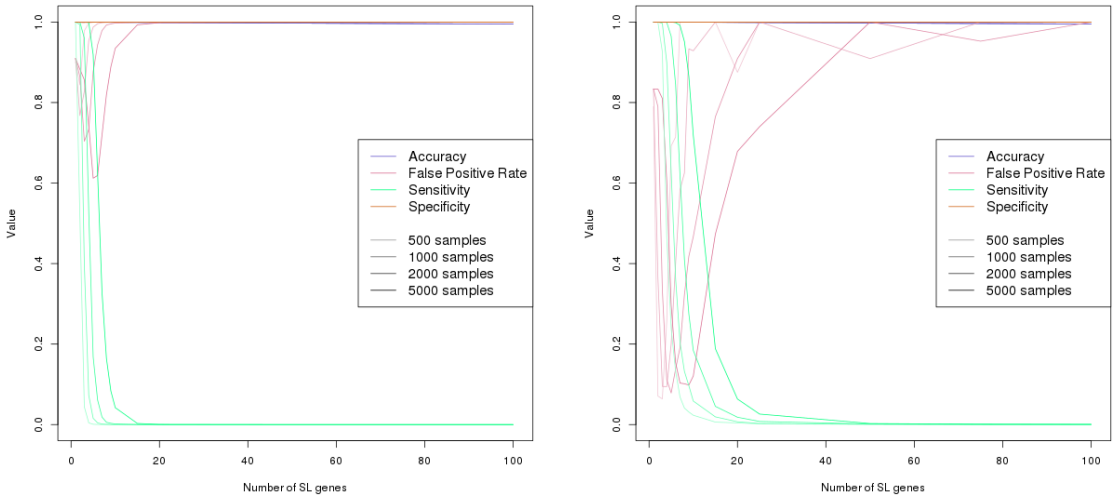


(b) ROC curves



(c) Statistical performance (AUROC)

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



(a) χ^2 testing without direction

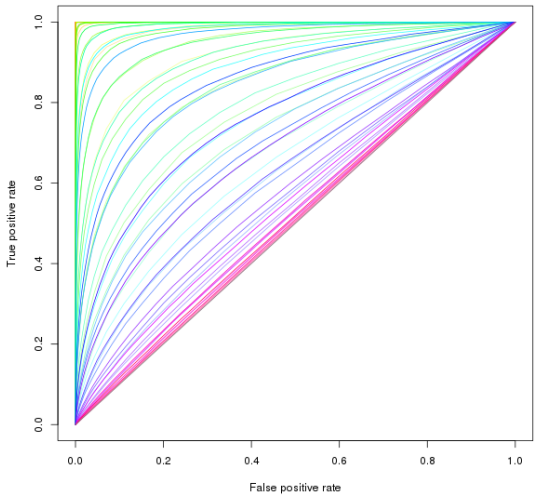
(b) SLIPT with directional criteria

Figure 3.18: **Statistical evaluation of directional criteria.** A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and the χ^2 test. SLIPT exhibited a consistently higher sensitivity and lower false positive rate.

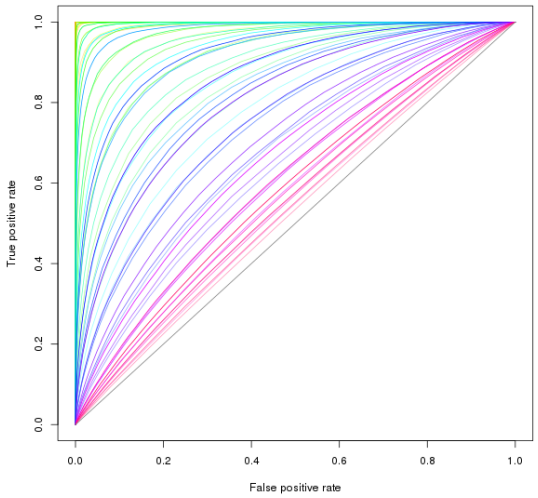
($r = 0.8$) with the query gene were added (as described in Section 3.3.2.1). These simulations had similar performance (in Figure 3.17) to those without these correlations with a higher specificity and a lower false positive rate (shown in Figure 3.17a).

The directional criteria of the SLIPT procedure was important in this case, enhancing its performance, particularly in distinguishing positively correlated non synthetic lethal genes. The multivariate normal simulations were performed, with 20,000 genes, including all of the correlation structures discussed (with synthetic lethal, non synthetic lethal, and query correlated genes). These simulations were compared for the direction SLIPT and the χ^2 testing. There was a considerably higher statistical performance with SLIPT, particularly increased sensitivity and lower false positive rate (as shown in Figure 3.18).

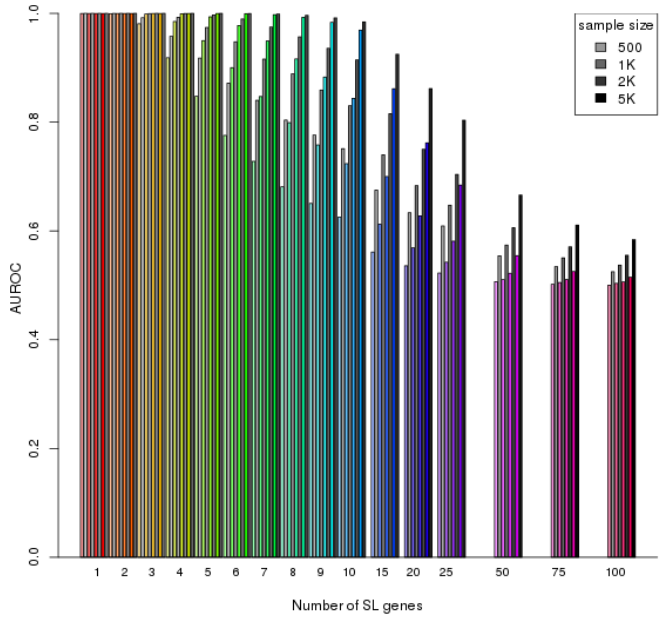
These results show that the performance of SLIPT is appropriate for the analysis of expression datasets, where positively correlated genes commonly occur, with the directional condition robustly improving the performance of SLIPT across simulation parameters (compared to the χ^2 test). Without assuming the underlying number of synthetic lethal genes, SLIPT will performed than the χ^2 test alone, irrespective of the significance threshold as shown by ROC analysis (in Figure 3.19). The directional



(a) χ^2 testing without direction



(b) SLIPT with directional criteria



(c) Statistical performance (AUROC)

Figure 3.19: **Performance with directional criteria.** A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT and χ^2 test. SLIPT had higher performance across simulation parameters, clearly differing from random (grey diagonal) in ROC curves up to 100 SL genes (b). The performance (c) of SLIPT (in greyscale) was consistently higher than the χ^2 test (in color).

SLIPT methodology outperformed the χ^2 test at detecting synthetic lethal partners with even up to 100 synthetic lethal genes.

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

3.4 Graph Structure Methods

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

3.4.1 Upstream and Downstream Gene Detection

Comparison of experimental and computational candidate synthetic lethal partner genes within pathway structures was performed to determine whether these sets of genes were related by pathway structure. Considering the differences in how these candidates were generated, it was unsurprising that they did not detect some identical genes within the candidate biological pathways. However, they could still be related by being upstream or downstream of each other.

Using the Reactome version 52 (Croft *et al.*, 2014), as described in Section 2.4.2, genes detected by each synthetic lethal discovery approach were mapped to the graph structure for each candidate pathway identified in Chapter 4 (with graphs defined as described in Section 2.4.3). To test whether siRNA candidate genes were upstream of SLIPT candidate genes, shortest paths were traced between each pair of these genes in a directed network. The paths where the siRNA candidate was upstream ("up") and downstream ("down") of a SLIPT candidate were scored. This procedure yielded the total number of shortest paths which indicated that siRNA genes were upstream or downstream of the SLIPT genes and measured the difference between these to determine if there was an imbalance in a particular direction. While this difference was indicative of the number of paths between the gene candidate groups in either direction, it was not sufficient to statistically verify structure or relationships between siRNA and SLIPT genes. It was be combined with a permutation resampling pro-

cedure (as described in Section 3.4.1.1) to test for directional relationships in either direction.

Initially, this procedure excluded genes that were detected by both approaches since they would count in both directions. Upon further consideration, these genes were restored to accounted for since they may contribute unequally to each gene set if there are unequal numbers of genes above or below them in the pathway structure.

3.4.1.1 Permutation Analysis for Statistical Significance

A permutation procedure was developed to randomly assign members of the pathway to siRNA and/or SLIPT groups, with the same number of each candidate partner gene set as observed in the pathway. These permuted genes were measured for pathway structure between the permuted gene groups as performed for the observed candidates (as performed in Section 3.4.1). A distribution of pathway structure relationships expected by chance was generated by permuting iteratively over these pathways. The resulting null distribution was compared to the observed counts of relationships (in either direction). This procedure yielded a permutation p-value as the proportion of permutations in which had a value greater than the observed value. The null hypothesis was that there was no relationship between these gene groups compared to genes that had been selected at random. Thus both the alternate hypotheses that the siRNA genes were either upstream of the SLIPT genes or that they are downstream of them were testable.

The permutation procedure does not assume the underlying distribution of the data under the null hypothesis and accounts for the total number of nodes, edges, siRNA, and SLIPT genes in each graph or networkpathway structure. The number of genes detected by both siRNA and SLIPT was not accounted for under the initial shortest path counts procedure that excluded them. Once they were included, it was ensured that the number of intersecting genes was equal to the number observed to test for pathway structure without changing the intersection size, the subject of prior analyses.

3.4.1.2 Hierarchy Based on Biological Context

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

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

$$hierarchy(z) = 1 + diameter(G).$$

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

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

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

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

This process could be performed iteratively to fill in pathway hierarchy but it was not necessary to perform further iterations for the candidate synthetic lethal pathways investigated which exhibited strong directional structure and the small world property (i.e., had a low diameter). Using this procedure, genes in a pathway graph structure were assigned to an integer valued hierarchy upstream to downstream by this procedure:

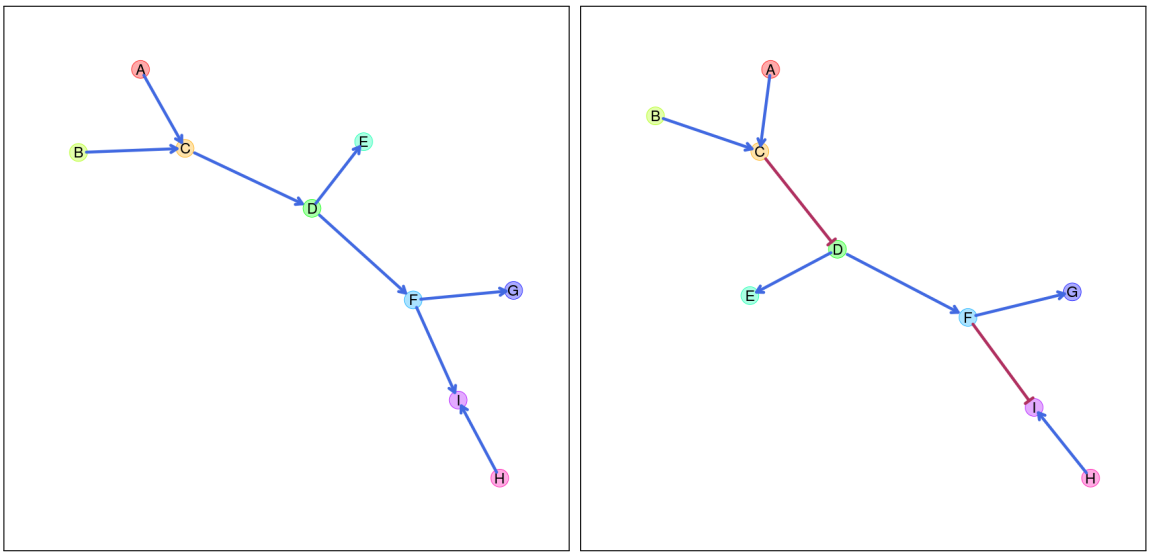
$$hierarchy \in \{1, 2, 3, ..., 1 + diameter(G)\}$$

This hierarchy of pathway directionality (e.g., that shown in Figure 5.7) was used for comparison with measures of the number of synthetic lethal partners detected by either approach.

3.4.2 Simulating Gene Expression from Graph Structures

The simulation procedure was refined to generate expression data with correlation structure from a known graph structure. This enabled modelling of synthetic lethal partners within a biological pathway and the investigation of the impact of pathway structure on synthetic lethal prediction. Firstly, a simulated pathway was constructed as a graph structure, with the `igraph` R package Csardi and Nepusz (2006), with the state of the edges (i.e, whether they activate or inhibit downstream pathway members). This simulation procedure was intended for biological pathway members with correlated

gene expression (higher than the background of genes in other pathways) but it may also be applicable to modelling protein levels (e.g, in a kinase regulation cascade) or substrates and products (e.g., in a metabolic pathway).

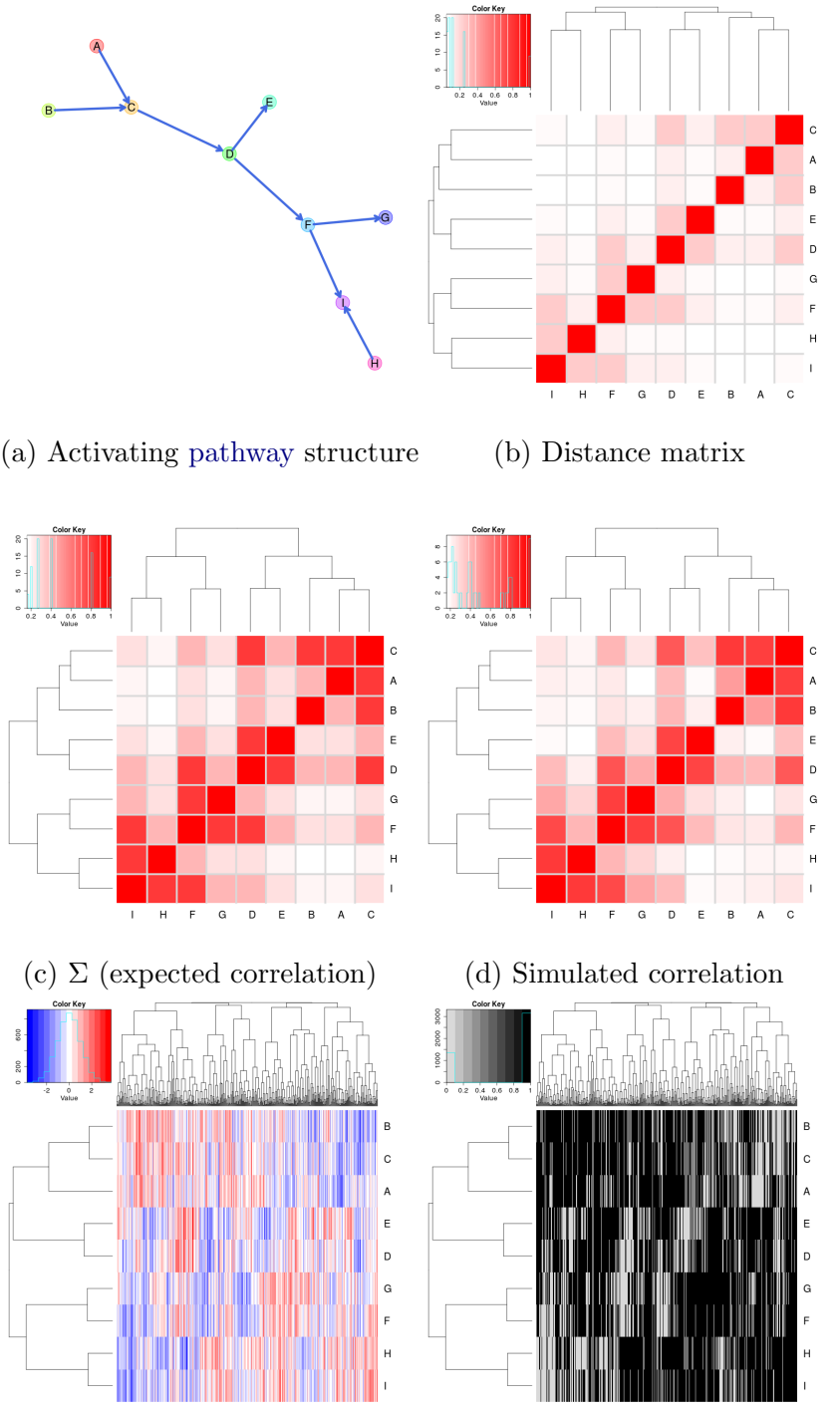


(a) Activating pathway structure (b) Pathway structure with inhibitions

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

The graph structure was constructed from which simulated data will be generated from, sampling a multivariate normal distribution using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016). Throughout this section, the simulation procedure will be demonstrated with the relatively simple constructed graph structure shown in Figure 3.20. This graph structure visualisation was specifically developed for (directed) iGraph objects in R and has been released in the `plot.igraph` package and `igraph.extensions` library (see Table 2.6 and Section 3.5.3). The `plot_directed` function enabled customisation of plot parameters for each node or edge and mixed (directed) edge types for indicating activation or inhibition. These inhibition links (which occur frequently in biological pathways) were demonstrated in Figure 3.20b.

The simulation procedure was designed to use such graph structures to inform development of a "Sigma" variance-covariance matrix (Σ) for sampling from a multivariate normal distribution (using the `mvtnorm` R package). Given a graph structure (or adjacency matrix), such as Figure 3.21a, a relation matrix was calculated based on distance such that nearer nodes are given higher weight than farther nodes. Through-

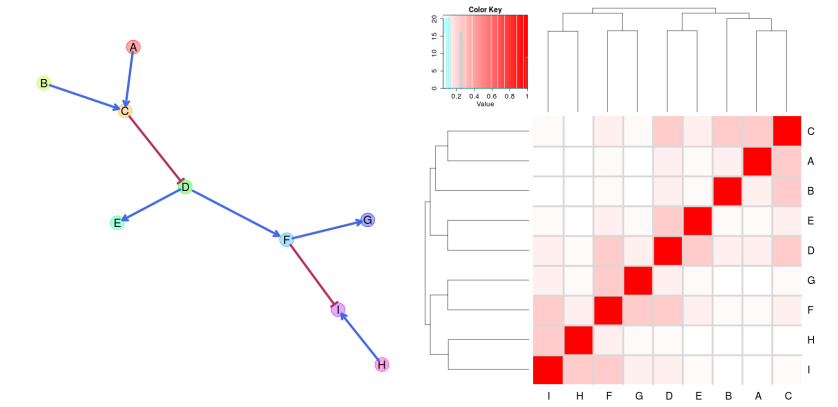


(a) Activating pathway structure (b) Distance matrix

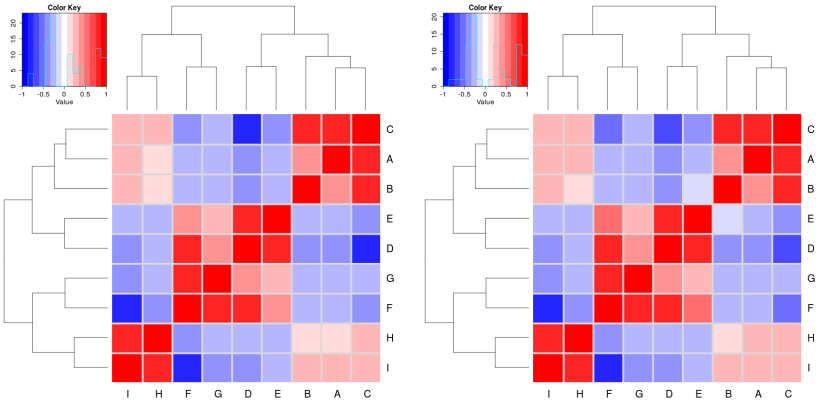
(c) Σ (expected correlation) (d) Simulated correlation

(e) Simulated expression data (f) Simulated gene function calls

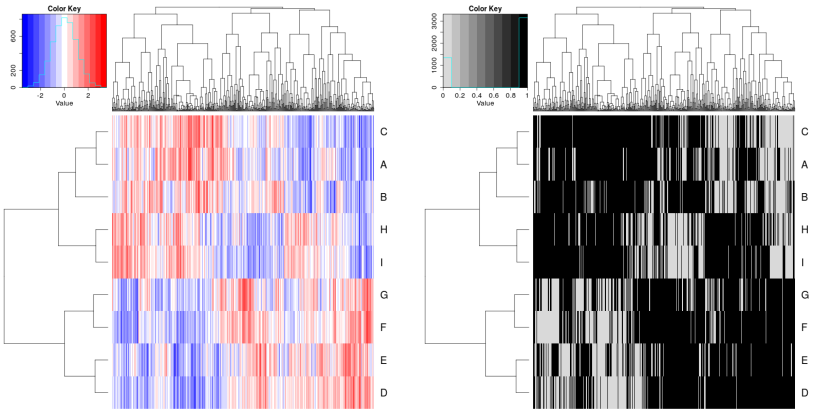
Figure 3.21: **Simulating expression from a graph structure.** An example graph structure that was used to derive a correlation structure from the relative distances between nodes and simulate continuous gene expression with sampling from the multivariate normal distribution.



(a) Inhibiting pathway structure (b) Distance matrix



(c) Σ (expected correlation) (d) Simulated correlation



(e) Simulated expression data (f) Simulated gene function calls

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

out this thesis, a geometrically decreasing (relative) distance weighting was used, with each more distant node being related by 1/2 compared to the next nearest, as shown in Figure 3.21b. An arithmetically decreasing (absolute) distance weighting is also supported in the `graphsim` R package release of this procedure.

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

The simulation procedure (depicted in Figure 3.21) can be used for pathways containing inhibition links (as shown in Figure 3.22) with several refinements. With the inhibition links (as shown in Figure 3.22a), distances were calculated in the same manner as before (Figure 3.22b) with inhibitions accounted for by iteratively multiplying downstream nodes by -1 to form blocks of negative correlations (as shown in Figures 3.22c and 3.22d). A multivariate normal distribution with these negative correlations can be sampled to generate simulated data (as shown in Figures 3.22e and 3.22f).

These simulated datasets could then be used for simulating synthetic lethal partners of a query gene within a graph network. The query gene was assumed to be separate from the graph network pathway and was added to the dataset using the procedure

in Section 3.2.2. Thus we can simulate known synthetic lethal partner genes within a synthetic lethal partner pathway structure.

3.5 Customised Functions and Packages Developed

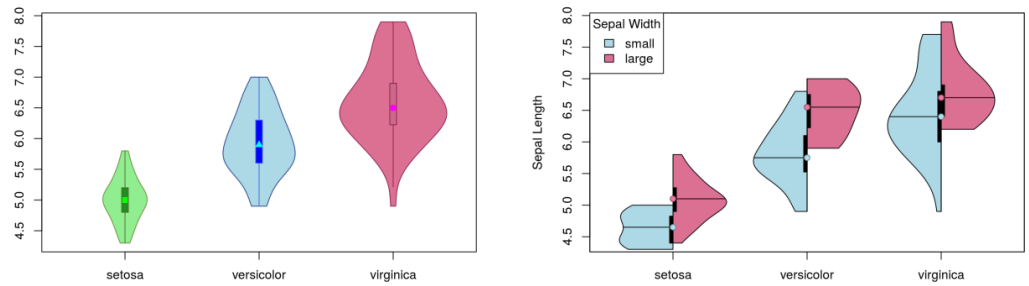
Various R packages (R Core Team, 2016) have been developed throughout this thesis using `devtools` (Wickham and Chang, 2016) and `roxygen` (Wickham *et al.*, 2017) to enable reproducibility of customised analysis and visualisation. Many of these have been documented, demonstrated in vignettes, and released on GitHub (`https://github.com/TomKellyGenetics`) to enable the research community to utilise them in their own analysis. These are summarised in Table 2.6, with the corresponding urls for their GitHub repository which contains instructions for installation with the `devtools` R package (Wickham and Chang, 2016) and links the relevant vignette(s).

3.5.1 Synthetic Lethal Interaction Prediction Tool

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

SLIPT can be used amenable to analysis of any effectively continuous measure of gene activity (e.g., microarray, RNA-Seq, protein abundance, or pathway metagenes). Executing `slipt` is straightforward: the `prep_data_for_SL` function scores samples as "low", "medium", or "high" for each gene, then the `detect_SL` function tests a given query gene against all potential partners by performing the chi-squared test and directional conditions. This function returns a table summarising the observed and expected sample numbers used for the directional criteria, the χ^2 values, and corresponding p-values including adjusting for multiple comparisons. The `count_of_SL` and `table_of_SL` functions serve to facilitate summary and extraction of the positive SLIPT hits, respectively, from the table of predictions of synthetic lethal partners.

The SLIPT methodology in this package release was used in later analyses rather than the corresponding source R code, including use on remote machines and upon simulated data. In particular, the functions in the package facilitate alterations to parameters, such as the proportion of samples called as exhibiting low or high gene activity (as shown in Section 6.1.1). This release supports reproducible research and enables wider use of SLIPT in future investigations into other disease genes.



(a) Customised violin plot

(b) Split violin plot

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

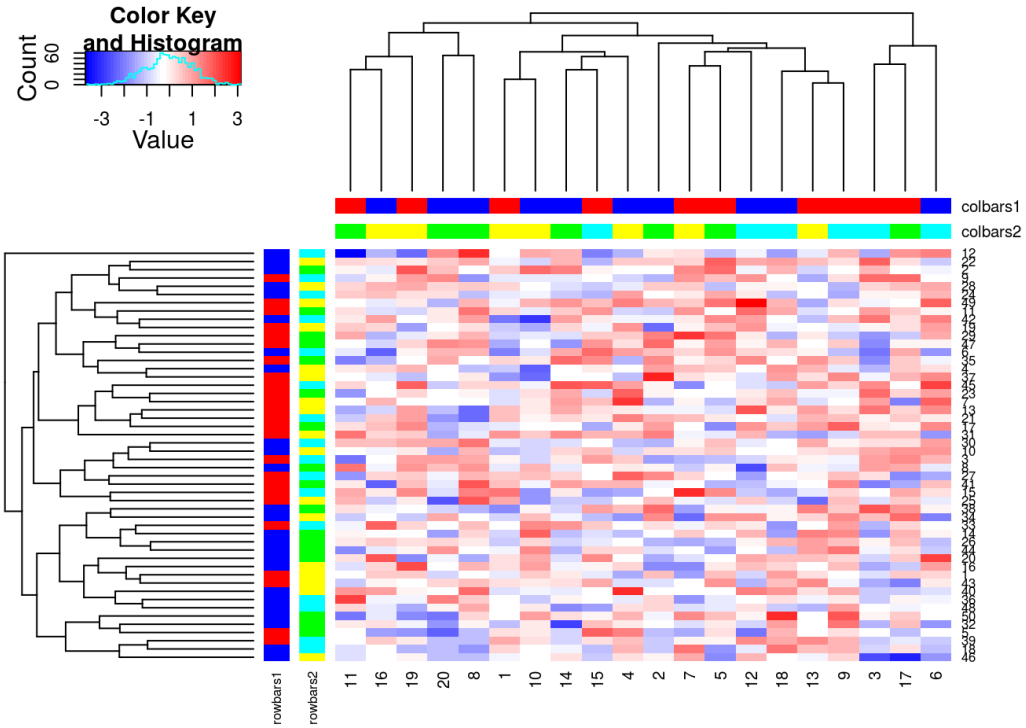


Figure 3.24: **Demonstration of annotated heatmap**. The example heatmap depicts the additional row and column annotation bars enabled by `heatmap.2x`, extending the features of `gplots` with backwards compatible inputs.

3.5.2 Data Visualisation

Customised data visualisations in R (R Core Team, 2016) were developed to present data throughout this thesis. The `vioplotx` and `heatmap.2x` packages are enhancements of the `vioplot` package (Adler, 2005) and `heatmap.2` provided by the `gplots` package (Warnes *et al.*, 2015).

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

The `heatmap.2x` function provides extensions for annotation colour bars for both the rows and columns (as shown in Figure 3.24). Multiple bars are enabled on both axes with matrix inputs (rather than single vector for `heatmap.2`) which facilitates additional plotting of gene and sample characteristics for comparison with correlation matrices, expression profiles, or pathway metagenes. The annotation bar inputs correspond to their orientation on the plot, each colour bar is provided as a column for the row annotation on the left of the heatmap and as a row for the column annotation on top of the heatmap. Row and column annotation bars are labelled with the column or row names respectively. Additional parameters enable resizing of these annotation bar labels and control of reordering columns for when samples have been ordered in advance (e.g., ranked by a metagene or split into groups clustered separately). These features were used through this thesis and have been provided in a package GitHub repository (`https://github.com/TomKellyGenetics/heatmap.2x`).

3.5.3 Extensions to the iGraph Package

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

gether as a collection of extensions to the `igraph` package (https://github.com/TomKellyGenetics/igraph.extensions).

3.5.3.1 Sampling Simulated Data from Graph Structures

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

3.5.3.2 Plotting Directed Graph Structures

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

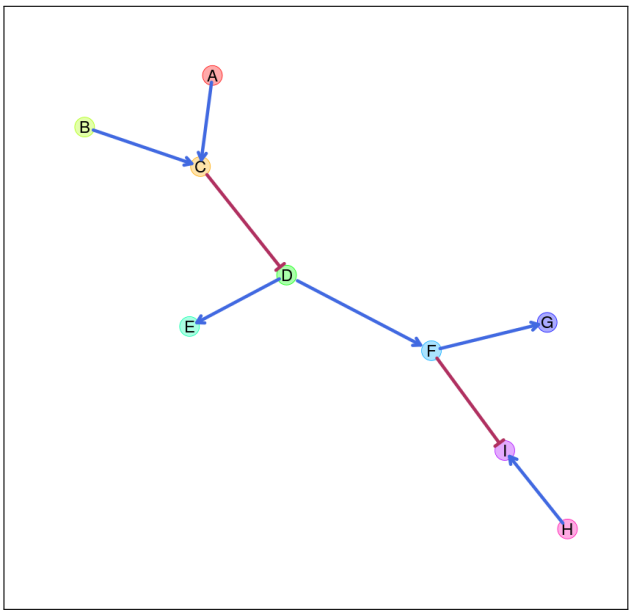


Figure 3.25: **Simulating graph structures.** An example graph structure which has been used throughout demonstrating the simulation procedure from graph structures. Activating links are denoted by blue arrows and inhibiting links by red edges.

3.5.3.3 Computing Information Centrality

The shortest paths of a network were computed by the `igraph` package (Csardi and Nepusz, 2006) which can be extended to calculate the network efficiency but was not provided by the package itself (ss described in Section 2.4.4). The "information centrality" of a vertex is computed as the relative change in the network efficiency when the vertex is removed. Information centrality is calculated iteratively for each node and the sum of information centrality for each vertex is the information centrality for the network. These metrics were released in the `info.centrality` package (`https://github.com/TomKellyGenetics/info.centrality`).

3.5.3.4 Testing Pathway Structure with Permutation Testing

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

3.5.3.5 Metapackage to Install iGraph Functions

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

Bibliography

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C.D., Annala, M., Aprikian, A., Armenia, J., Arora, A., *et al.* (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, **163**(4): 1011–1025.

Adler, D. (2005) *vioplot: Violin plot.* R package version 0.2.

Akbani, R., Akdemir, K.C., Aksoy, B.A., Albert, M., Ally, A., Amin, S.B., Arachchi, H., Arora, A., Auman, J.T., Ayala, B., *et al.* (2015) Genomic Classification of Cutaneous Melanoma. *Cell*, **161**(7): 1681–1696.

Akobeng, A.K. (2007) Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Pdiatrica*, **96**(5): 644–647.

American Cancer Society (2017) Genetics and cancer. https://www.cancer.org/cancer/cancer-causes/genetics.html. Accessed: 22/03/2017.

Anjomshoaa, A., Lin, Y.H., Black, M.A., McCall, J.L., Humar, B., Song, S., Fukuzawa, R., Yoon, H.S., Holzmann, B., Friederichs, J., *et al.* (2008) Reduced expression of a gene proliferation signature is associated with enhanced malignancy in colon cancer. *Br J Cancer*, **99**(6): 966–973.

Araki, H., Knapp, C., Tsai, P., and Print, C. (2012) GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**: 76–82.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1): 25–29.

Ashworth, A. (2008) A synthetic lethal therapeutic approach: poly(adp) ribose polymerase inhibitors for the treatment of cancers deficient in dna double-strand break repair. *J Clin Oncol*, **26**(22): 3785–90.

Ashworth, A., Lord, C.J., and Reis-Filho, J.S. (2011) Genetic interactions in cancer progression and treatment. *Cell*, **145**(1): 30–38.

Audeh, M.W., Carmichael, J., Penson, R.T., Friedlander, M., Powell, B., Bell-McGuinn, K.M., Scott, C., Weitzel, J.N., Oaknin, A., Loman, N., *et al.* (2010) Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 245–51.

Babyak, M.A. (2004) What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*, **66**(3): 411–21.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R., *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*, **91**(2): 355–358.

Barabási, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439): 509–12.

Barabási, A.L., Gulbahce, N., and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**(1): 56–68.

Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**(2): 101–13.

Barrat, A. and Weigt, M. (2000) On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems*, **13**(3): 547–560.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391): 603–607.

Barry, W.T. (2016) *safe: Significance Analysis of Function and Expression*. R package version 3.14.0.

Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B., and Boone, C. (2010a) Synthetic genetic array (sga) analysis in saccharomyces cerevisiae and schizosaccharomyces pombe. *Methods Enzymol*, **470**: 145–79.

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., *et al.* (2010b) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Meth*, **7**(12): 1017–1024.

Bass, A.J., Thorsson, V., Shmulevich, I., Reynolds, S.M., Miller, M., Bernard, B., Hinoue, T., Laird, P.W., Curtis, C., Shen, H., *et al.* (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**(7517): 202–209.

Bates, D. and Maechler, M. (2016) *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-7.1.

Bateson, W. and Mendel, G. (1909) *Mendel's principles of heredity, by W. Bateson*. University Press, Cambridge [Eng.].

Becker, K.F., Atkinson, M.J., Reich, U., Becker, I., Nekarda, H., Siewert, J.R., and Hfler, H. (1994) E-cadherin gene mutations provide clues to diffuse type gastric carcinomas. *Cancer Research*, **54**(14): 3845–3852.

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., *et al.* (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353): 609–615.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**(1): 289–300.

Berx, G., Cleton-Jansen, A.M., Nollet, F., de Leeuw, W.J., van de Vijver, M., Cornelisse, C., and van Roy, F. (1995) E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. *EMBO J*, **14**(24): 6107–15.

Berx, G., Cleton-Jansen, A.M., Strumane, K., de Leeuw, W.J., Nollet, F., van Roy, F., and Cornelisse, C. (1996) E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain. *Oncogene*, **13**(9): 1919–25.

Berx, G. and van Roy, F. (2009) Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harb Perspect Biol*, **1**: a003129.

Bitler, B.G., Aird, K.M., Garipov, A., Li, H., Amatangelo, M., Kossenkov, A.V., Schultz, D.C., Liu, Q., Shih Ie, M., Conejo-Garcia, J.R., *et al.* (2015) Synthetic lethality by targeting ezh2 methyltransferase activity in arid1a-mutated cancers. *Nat Med*, **21**(3): 231–8.

Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Burgess, S., Buza, T., Gresham, C., *et al.* (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res*, **43**(Database issue): D1049–1056.

Boone, C., Bussey, H., and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, **8**(6): 437–49.

Borgatti, S.P. (2005) Centrality and network flow. *Social Networks*, **27**(1): 55 – 71.

Boucher, B. and Jenna, S. (2013) Genetic interaction networks: better understand to better predict. *Front Genet*, **4**: 290.

Bozovic-Spasojevic, I., Azambuja, E., McCaskill-Stevens, W., Dinh, P., and Cardoso, F. (2012) Chemoprevention for breast cancer. *Cancer treatment reviews*, **38**(5): 329–339.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**(1): 5–32.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, **30**(1): 107 – 117.

Brouxhon, S.M., Kyrkanides, S., Teng, X., Athar, M., Ghazizadeh, S., Simon, M., O'Banion, M.K., and Ma, L. (2014) Soluble E-cadherin: a critical oncogene modulating receptor tyrosine kinases, MAPK and PI3K/Akt/mTOR signaling. *Oncogene*, **33**(2): 225–235.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009) Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*, **10**(6): 2763–2788.

Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., and Helleday, T. (2005) Specific killing of *BRCA2*-deficient tumours with inhibitors of poly*adpribose* polymerase. *Nature*, **434**(7035): 913–7.

Bussey, H., Andrews, B., and Boone, C. (2006) From worm genetic networks to complex human diseases. *Nat Genet*, **38**(8): 862–3.

Butland, G., Babu, M., Diaz-Mejia, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O., *et al.* (2008) esga: E. coli synthetic genetic array analysis. *Nat Methods*, **5**(9): 789–95.

cBioPortal for Cancer Genomics (cBioPortal) (2017) cBioPortal for Cancer Genomics. http://www.cbioportal.org/. Accessed: 26/03/2017.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, **39**(Database issue): D685–690.

Chen, A., Beetham, H., Black, M.A., Priya, R., Telford, B.J., Guest, J., Wiggins, G.A.R., Godwin, T.D., Yap, A.S., and Guilford, P.J. (2014) E-cadherin loss alters cytoskeletal organization and adhesion in non-malignant breast cells but is insufficient to induce an epithelial-mesenchymal transition. *BMC Cancer*, **14**(1): 552.

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

Chipman, K. and Singh, A. (2009) Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, **10**(1): 17.

Christofori, G. and Semb, H. (1999) The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends in Biochemical Sciences*, **24**(2): 73 – 76.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., *et al.* (2015) Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, **163**(2): 506–519.

Clark, M.J. (2004) Endogenous Regulator of G Protein Signaling Proteins Suppress G o-Dependent -Opioid Agonist-Mediated Adenylyl Cyclase Supersensitization. *Journal of Pharmacology and Experimental Therapeutics*, **310**(1): 215–222.

Collingridge, D.S. (2013) A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, **7**(1): 81–97.

Collins, F.S. and Barker, A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am*, **296**(3): 50–57.

Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer, D., Cope, L., Creighton, C., Danilova, L., *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**(7511): 543–550.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., *et al.* (2010) The genetic landscape of a cell. *Science*, **327**(5964): 425–31.

Costanzo, M., Baryshnikova, A., Myers, C.L., Andrews, B., and Boone, C. (2011) Charting the genetic interaction map of a cell. *Curr Opin Biotechnol*, **22**(1): 66–74.

Courtney, K.D., Corcoran, R.B., and Engelman, J.A. (2010) The PI3K pathway as drug target in human cancer. *J Clin Oncol*, **28**(6): 1075–1083.

Creighton, C.J., Morgan, M., Gunaratne, P.H., Wheeler, D.A., Gibbs, R.A., Robertson, A., Chu, A., Beroukhim, R., Cibulskis, K., Signoretti, S., *et al.* (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, **499**(7456): 43–49.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.* (2014) The Reactome pathway 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.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res*, **5**(10): 2929–2943.

Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y., *et al.* (2005) The synthetic genetic interaction spectrum of essential genes. *Nat Genet*, **37**(10): 1147–1152.

De Leeuw, W.J., Berx, G., Vos, C.B., Peterse, J.L., Van de Vijver, M.J., Litvinov, S., Van Roy, F., Cornelisse, C.J., and Cleton-Jansen, A.M. (1997) Simultaneous loss of E-cadherin and catenins in invasive lobular breast cancer and lobular carcinoma in situ. *J Pathol*, **183**(4): 404–11.

De Santis, G., Miotti, S., Mazzi, M., Canevari, S., and Tomassetti, A. (2009) E-cadherin directly contributes to PI3K/AKT activation by engaging the PI3K-p85 regulatory subunit to adherens junctions of ovarian carcinoma cells. *Oncogene*, **28**(9): 1206–1217.

Demir, E., Babur, O., Rodchenkov, I., Aksoy, B.A., Fukuda, K.I., Gross, B., Sumer, O.S., Bader, G.D., and Sander, C. (2013) Using biological pathway data with Paxtools. *PLoS Comput Biol*, **9**(9): e1003194.

Deshpande, R., Asiedu, M.K., Klebig, M., Sutor, S., Kuzmin, E., Nelson, J., Piotrowski, J., Shin, S.H., Yoshida, M., Costanzo, M., *et al.* (2013) A comparative genomic approach for identifying synthetic lethal interactions in human cancer. *Cancer Res*, **73**(20): 6128–36.

Dickson, D. (1999) Wellcome funds cancer database. *Nature*, **401**(6755): 729.

Dijkstra, E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**(1): 269–271.

Dixon, S.J., Andrews, B.J., and Boone, C. (2009) Exploring the conservation of synthetic lethal genetic interaction networks. *Commun Integr Biol*, **2**(2): 78–81.

Dixon, S.J., Fedyshyn, Y., Koh, J.L., Prasad, T.S., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.L., *et al.* (2008) Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A*, **105**(43): 16653–8.

Dong, L.L., Liu, L., Ma, C.H., Li, J.S., Du, C., Xu, S., Han, L.H., Li, L., and Wang, X.W. (2012) E-cadherin promotes proliferation of human ovarian cancer cells in vitro via activating MEK/ERK pathway. *Acta Pharmacol Sin*, **33**(6): 817–822.

Dorsam, R.T. and Gutkind, J.S. (2007) G-protein-coupled receptors and cancer. *Nat Rev Cancer*, **7**(2): 79–94.

Erdős, P. and Rényi, A. (1959) On random graphs I. *Publ Math Debrecen*, **6**: 290–297.

Erdős, P. and Rényi, A. (1960) On the evolution of random graphs. In *Publ. Math. Inst. Hung. Acad. Sci*, volume 5, 17–61.

Eroles, P., Bosch, A., Perez-Fidalgo, J.A., and Lluch, A. (2012) Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer Treat Rev*, **38**(6): 698–707.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., *et al.* (2005) Targeting the dna repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**(7035): 917–21.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8): 861 – 874. {ROC} Analysis in Pattern Recognition.

Fece de la Cruz, F., Gapp, B.V., and Nijman, S.M. (2015) Synthetic lethal vulnerabilities of cancer. *Annu Rev Pharmacol Toxicol*, **55**: 513–531.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, **136**(5): E359–386.

Fisher, R.A. (1919) Xv.the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52**(02): 399–433.

Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J., *et al.* (2009) Inhibition of poly(adp-ribose) polymerase in tumors from BRCA mutation carriers. *N Engl J Med*, **361**(2): 123–34.

Fong, P.C., Yap, T.A., Boss, D.S., Carden, C.P., Mergui-Roelvink, M., Gourley, C., De Greve, J., Lubinski, J., Shanley, S., Messiou, C., *et al.* (2010) Poly(adp)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. *J Clin Oncol*, **28**(15): 2512–9.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., *et al.* (2015) COSMIC: exploring the world's

knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**(Database issue): D805–811.

Fraser, A. (2004) Towards full employment: using RNAi to find roles for the redundant. *Oncogene*, **23**(51): 8346–52.

Fromental-Ramain, C., Warot, X., Lakkaraju, S., Favier, B., Haack, H., Birling, C., Dierich, A., Doll e, P., and Chambon, P. (1996) Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning. *Development*, **122**(2): 461–472.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004) A census of human cancer genes. *Nat Rev Cancer*, **4**(3): 177–183.

Futreal, P.A., Kasprzyk, A., Birney, E., Mullikin, J.C., Wooster, R., and Stratton, M.R. (2001) Cancer and genomics. *Nature*, **409**(6822): 850–852.

Gao, B. and Roux, P.P. (2015) Translational control by oncogenic signaling pathways. *Biochimica et Biophysica Acta*, **1849**(7): 753–65.

Gatza, M.L., Kung, H.N., Blackwell, K.L., Dewhirst, M.W., Marks, J.R., and Chi, J.T. (2011) Analysis of tumor environmental response and oncogenic pathway activation identifies distinct basal and luminal features in HER2-related breast tumor subtypes. *Breast Cancer Res*, **13**(3): R62.

Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A., *et al.* (2010) A pathway-based classification of human breast cancer. *Proc Natl Acad Sci USA*, **107**(15): 6994–6999.

Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C., and Perou, C.M. (2014) An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet*, **46**(10): 1051–1059.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10): R80.

Genz, A. and Bretz, F. (2009) Computation of multivariate normal and t probabilities. In *Lecture Notes in Statistics*, volume 195. Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2016) *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5. URL.

Glaire, M.A., Brown, M., Church, D.N., and Tomlinson, I. (2017) Cancer predisposition syndromes: lessons for truly precision medicine. *J Pathol*, **241**(2): 226–235.

Globus (Globus) (2017) Research data management simplified. https://www.globus.org/. Accessed: 25/03/2017.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**(6): 333–351.

Grady, W.M., Willis, J., Guilford, P.J., Dunbier, A.K., Toro, T.T., Lynch, H., Wiesner, G., Ferguson, K., Eng, C., Park, J.G., *et al.* (2000) Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat Genet*, **26**(1): 16–17.

Graziano, F., Humar, B., and Guilford, P. (2003) The role of the E-cadherin gene (*CDH1*) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice. *Annals of Oncology*, **14**(12): 1705–1713.

Guaragnella, N., Palermo, V., Galli, A., Moro, L., Mazzoni, C., and Giannattasio, S. (2014) The expanding role of yeast in cancer research and diagnosis: insights into the function of the oncosuppressors p53 and BRCA1/2. *FEMS Yeast Res*, **14**(1): 2–16.

Güell, O., Sagus, F., and Serrano, M. (2014) Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. *PLoS Comput Biol*, **10**(5): e1003637.

Guilford, P. (1999) E-cadherin downregulation in cancer: fuel on the fire? *Molecular Medicine Today*, **5**(4): 172 – 177.

Guilford, P., Hopkins, J., Harraway, J., McLeod, M., McLeod, N., Harawira, P., Taite, H., Scoular, R., Miller, A., and Reeve, A.E. (1998) E-cadherin germline mutations in familial gastric cancer. *Nature*, **392**(6674): 402–5.

Guilford, P., Humar, B., and Blair, V. (2010) Hereditary diffuse gastric cancer: translation of *CDH1* germline mutations into clinical practice. *Gastric Cancer*, **13**(1): 1–10.

Guilford, P.J., Hopkins, J.B., Grady, W.M., Markowitz, S.D., Willis, J., Lynch, H., Rajput, A., Wiesner, G.L., Lindor, N.M., Burgart, L.J., *et al.* (1999) E-cadherin germline mutations define an inherited cancer syndrome dominated by diffuse gastric cancer. *Hum Mutat*, **14**(3): 249–55.

Guo, J., Liu, H., and Zheng, J. (2016) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res*, **44**(D1): D1011–1017.

Hajian-Tilaki, K. (2013) Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, **4**(2): 627–635.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009) The weka data mining software: an update. *SIGKDD Explor Newsl*, **11**(1): 10–18.

Hammerman, P.S., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., *et al.* (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**(7417): 519–525.

Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**(1): 57–70.

Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**(5): 646–674.

Hanna, S. (2003) Cancer incidence in new zealand (2003-2007). In D. Forman, D. Bray F Brewster, C. Gombe Mbalawa, B. Kohler, M. Piñeros, E. Steliarova-Foucher, R. Swaminathan, and J. Ferlay (editors), *Cancer Incidence in Five Continents*, volume X, 902–907. International Agency for Research on Cancer, Lyon, France. Electronic version http://ci5.iarc.fr Accessed 22/03/2017.

Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., Schrader, K.A., Schaeffer, D.F., Shumansky, K., Zogopoulos, G., *et al.* (2015) Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol*, **1**(1): 23–32.

Heiskanen, M.A. and Aittokallio, T. (2012) Mining high-throughput screens for cancer drug targets-lessons from yeast chemical-genomic profiling and synthetic lethality.

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

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

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

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

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

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

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

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

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

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

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

International HapMap 3 Consortium (HapMap) (2003) The International HapMap Project. *Nature*, **426**(6968): 789–796.

Jeanes, A., Gottardi, C.J., and Yap, A.S. (2008) Cadherins and cancer: how does cadherin dysfunction promote tumor progression? *Oncogene*, **27**(55): 6920–6929.

Jerby-Arnon, L., Pfetzer, N., Waldman, Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P., *et al.* (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, **158**(5): 1199–1209.

Joachims, T. (1999) Making large-scale support vector machine learning practical. In S. Bernhard, lkopf, J.C.B. Christopher, and J.S. Alexander (editors), *Advances in kernel methods*, 169–184. MIT Press.

Ju, Z., Liu, W., Roebuck, P.L., Siwak, D.R., Zhang, N., Lu, Y., Davies, M.A., Akbani, R., Weinstein, J.N., Mills, G.B., *et al.* (2015) Development of a robust classifier for quality control of reverse-phase protein arrays. *Bioinformatics*, **31**(6): 912.

Kaelin, Jr, W. (2005) The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer*, **5**(9): 689–98.

Kaelin, Jr, W. (2009) Synthetic lethality: a framework for the development of wiser cancer therapeutics. *Genome Med*, **1**: 99.

Kamada, T. and Kawai, S. (1989) An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**(1): 7–15.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**(6821): 685–690.

Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotech*, **23**(5): 561–566.

Kelly, S.T. (2013) *Statistical Predictions of Synthetic Lethal Interactions in Cancer*. Dissertation, University of Otago.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009)

Human Protein Reference Database–2009 update. *Nucleic Acids Res*, **37**(Database issue): D767–772.

Kim, N.G., Koh, E., Chen, X., and Gumbiner, B.M. (2011) E-cadherin mediates contact inhibition of proliferation through Hippo signaling-pathway components. *Proc Natl Acad Sci USA*, **108**(29): 11930–11935.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418): 61–70.

Kockel, L., Zeitlinger, J., Staszewski, L.M., Mlodzik, M., and Bohmann, D. (1997) Jun in drosophila development: redundant and nonredundant functions and regulation by two mapk signal transduction pathways. *Genes & Development*, **11**(13): 1748–1758.

Kozlov, K.N., Gursky, V.V., Kulakovskiy, I.V., and Samsonova, M.G. (2015) Sequence-based model of gap gene regulation network. *BMC Genomics*, **15**(Suppl 12): S6.

Kranthi, S., Rao, S., and Manimaran, P. (2013) Identification of synthetic lethal pairs in biological systems through network information centrality. *Mol BioSyst*, **9**(8): 2163–2167.

Kroepil, F., Fluegen, G., Totikov, Z., Baldus, S.E., Vay, C., Schauer, M., Topp, S.A., Esch, J.S., Knoefel, W.T., and Stoecklein, N.H. (2012) Down-regulation of CDH1 is associated with expression of SNAI1 in colorectal adenomas. *PLoS ONE*, **7**(9): e46665.

Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**(7333): 187–197.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**(6822): 860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3): R25.

Latora, V. and Marchiori, M. (2001) Efficient behavior of small-world networks. *Phys Rev Lett*, **87**: 198701.

Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013) Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat Methods*, **10**(5): 427–31.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, **15**(2): R29.

Le Meur, N. and Gentleman, R. (2008) Modeling synthetic lethality. *Genome Biol*, **9**(9): R135.

Le Meur, N., Jiang, Z., Liu, T., Mar, J., and Gentleman, R.C. (2014) Slgi: Synthetic lethal genetic interaction. r package version 1.26.0.

Lee, A.Y., Perreault, R., Harel, S., 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, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**(4): 493–500.

Li, X.J., Mishra, S.K., Wu, M., Zhang, F., and Zheng, J. (2014) Syn-lethality: An integrative knowledge base of synthetic lethality towards discovery of selective 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.

Markowetz, F. (2017) All biology is computational biology. *PLoS Biol*, **15**(3): e2002050.

Masciari, S., Larsson, N., Senz, J., Boyd, N., Kaurah, P., Kandel, M.J., Harris, L.N., Pinheiro, H.C., Troussard, A., Miron, P., *et al.* (2007) Germline E-cadherin mutations in familial lobular breast cancer. *J Med Genet*, **44**(11): 726–31.

Mattison, J., van der Weyden, L., Hubbard, T., and Adams, D.J. (2009) Cancer gene discovery in mouse and man. *Biochim Biophys Acta*, **1796**(2): 140–161.

McLachlan, J., George, A., and Banerjee, S. (2016) The current status of parp inhibitors in ovarian cancer. *Tumori*, **102**(5): 433–440.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., Mastrogianakis, G.M., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216): 1061–1068.

Miles, D.W. (2001) Update on HER-2 as a target for cancer therapy: herceptin in the clinical setting. *Breast Cancer Res*, **3**(6): 380–384.

Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., *et al.* (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**(7407): 330–337.

Nagalla, S., Chou, J.W., Willingham, M.C., Ruiz, J., Vaughn, J.P., Dubey, P., Lash, T.L., Hamilton-Dutoit, S.J., Bergh, J., Sotiriou, C., *et al.* (2013) Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol*, **14**(4): R34.

Neeley, E.S., Kornblau, S.M., Coombes, K.R., and Baggerly, K.A. (2009) Variable slope normalization of reverse phase protein arrays. *Bioinformatics*, **25**(11): 1384.

Novomestky, F. (2012) *matrixcalc: Collection of functions for matrix calculations*. R package version 1.0-3.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997) Evolution of genetic redundancy. *Nature*, **388**(6638): 167–171.

Oliveira, C., Senz, J., Kaurah, P., Pinheiro, H., Sanges, R., Haegert, A., Corso, G., Schouten, J., Fitzgerald, R., Vogelsang, H., *et al.* (2009) Germline *CDH1* deletions in hereditary diffuse gastric cancer families. *Human Molecular Genetics*, **18**(9): 1545–1555.

Oliveira, C., Seruca, R., Hoogerbrugge, N., Ligtenberg, M., and Carneiro, F. (2013) Clinical utility gene card for: Hereditary diffuse gastric cancer (HDGC). *Eur J Hum Genet*, **21**(8).

Pandey, G., Zhang, B., Chang, A.N., Myers, C.L., Zhu, J., Kumar, V., and Schadt, E.E. (2010) An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol*, **6**(9).

Parker, J., Mullins, M., Cheung, M., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8): 1160–1167.

Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., *et al.* (2016) Erratum: The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*, **7**: 11908.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**(6797): 747–752.

Polyak, K. and Weinberg, R.A. (2009) Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nat Rev Cancer*, **9**(4): 265–73.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. R version 3.3.2.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7): e47.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., *et al.* (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, **322**(5900): 405–10.

Roychowdhury, S. and Chinnaiyan, A.M. (2016) Translating cancer genomes and transcriptomes for precision oncology. *CA Cancer J Clin*, **66**(1): 75–88.

Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. *Nat Rev Genet*, **14**(2): 89–99.

Ryan, C., Lord, C., and Ashworth, A. (2014) Daisy: Picking synthetic lethals from cancer genomes. *Cancer Cell*, **26**(3): 306–308.

Schena, M. (1996) Genome analysis with gene expression microarrays. *Bioessays*, **18**(5): 427–431.

Scheuer, L., Kauff, N., Robson, M., Kelly, B., Barakat, R., Satagopan, J., Ellis, N., Hensley, M., Boyd, J., Borgen, P., *et al.* (2002) Outcome of preventive surgery and screening for breast and ovarian cancer in BRCA mutation carriers. *J Clin Oncol*, **20**(5): 1260–1268.

Semb, H. and Christofori, G. (1998) The tumor-suppressor function of E-cadherin. *Am J Hum Genet*, **63**(6): 1588–93.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) Rocr: visualizing classifier performance in r. *Bioinformatics*, **21**(20): 7881.

Slurm development team (Slurm) (2017) Slurm workload manager. https://slurm.schedmd.com/. Accessed: 25/03/2017.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*, **98**(19): 10869–10874.

Srihari, S., Singla, J., Wong, L., and Ragan, M.A. (2015) Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biology Direct*, **10**(1): 57.

Stajich, J.E. and Lapp, H. (2006) Open source tools and toolkits for bioinformatics: significance, and where are we? *Brief Bioinformatics*, **7**(3): 287–296.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009) The cancer genome. *Nature*, **458**(7239): 719–724.

Ström, C. and Helleday, T. (2012) Strategies for the use of poly(adenosine diphosphate ribose) polymerase (parp) inhibitors in cancer therapy. *Biomolecules*, **2**(4): 635–649.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res*, **21**(12): 2213–2223.

Telford, B.J., Chen, A., Beetham, H., Frick, J., Brew, T.P., Gould, C.M., Single, A., Godwin, T., Simpson, K.J., and Guilford, P. (2015) Synthetic lethal screens identify

vulnerabilities in gpcr signalling and cytoskeletal organization in E-cadherin-deficient cells. *Mol Cancer Ther*, **14**(5): 1213–1223.

The 1000 Genomes Project Consortium (1000 Genomes) (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319): 1061–1073.

The Cancer Genome Atlas Research Network (TCGA) (2017) The Cancer Genome Atlas Project. https://cancergenome.nih.gov/. Accessed: 26/03/2017.

The Catalogue Of Somatic Mutations In Cancer (COSMIC) (2016) Cosmic: The catalogue of somatic mutations in cancer. http://cancer.sanger.ac.uk/cosmic. Release 79 (23/08/2016), Accessed: 05/02/2017.

The Comprehensive R Archive Network (CRAN) (2017) Cran. https://cran.r-project.org/. Accessed: 24/03/2017.

The ENCODE Project Consortium (ENCODE) (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**(5696): 636–640.

The National Cancer Institute (NCI) (2015) The genetics of cancer. https://www.cancer.gov/about-cancer/causes-prevention/genetics. Published: 22/04/2015, Accessed: 22/03/2017.

The New Zealand eScience Infrastructure (NeSI) (2017) NeSI. https://www.nesi.org.nz/. Accessed: 25/03/2017.

Tierney, L., Rossini, A.J., Li, N., and Sevcikova, H. (2015) *snow: Simple Network of Workstations*. R package version 0.4-2.

Tiong, K.L., Chang, K.C., Yeh, K.T., Liu, T.Y., Wu, J.H., Hsieh, P.H., Lin, S.H., Lai, W.Y., Hsu, Y.C., Chen, J.Y., *et al.* (2014) Csnk1e/ctnnb1 are synthetic lethal to tp53 in colorectal cancer and are markers for prognosis. *Neoplasia*, **16**(5): 441–50.

Tischler, J., Lehner, B., and Fraser, A.G. (2008) Evolutionary plasticity of genetic interaction networks. *Nat Genet*, **40**(4): 390–391.

Tomasetti, C. and Vogelstein, B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**(6217): 78–81.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**(5550): 2364–8.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**(5659): 808–13.

Tran, B., Dancey, J.E., Kamel-Reid, S., McPherson, J.D., Bedard, P.L., Brown, A.M., Zhang, T., Shaw, P., Onetto, N., Stein, L., *et al.* (2012) Cancer genomics: technology, discovery, and translation. *J Clin Oncol*, **30**(6): 647–660.

Travers, J. and Milgram, S. (1969) An experimental study of the small world problem. *Sociometry*, **32**(4): 425–443.

Tunggal, J.A., Helfrich, I., Schmitz, A., Schwarz, H., Gunzel, D., Fromm, M., Kemler, R., Krieg, T., and Niessen, C.M. (2005) E-cadherin is essential for in vivo epidermal barrier function by regulating tight junctions. *EMBO J*, **24**(6): 1146–1156.

Tutt, A., Robson, M., Garber, J.E., Domchek, S.M., Audeh, M.W., Weitzel, J.N., Friedlander, M., Arun, B., Loman, N., Schmutzler, R.K., *et al.* (2010) Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with *BRCA1* or *BRCA2* mutations and advanced breast cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 235–44.

University of California, Santa Cruz (UCSC) (2012) Ucsc cancer browser. Accessed 29/03/2012.

van der Post, R.S., Vogelaar, I.P., Carneiro, F., Guilford, P., Huntsman, D., Hoogerbrugge, N., Caldas, C., Schreiber, K.E., Hardwick, R.H., Ausems, M.G., *et al.* (2015) Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. *J Med Genet*, **52**(6): 361–374.

van Steen, K. (2012) Travelling the world of 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.

Vizeacoumar, F.J., Arnold, R., Vizeacoumar, F.S., Chandrashekhar, M., Buzina, A., Young, J.T., Kwan, J.H., Sayad, A., Mero, P., Lawo, S., *et al.* (2013) A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Mol Syst Biol*, **9**: 696.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**(6127): 1546–1558.

Vos, C.B., Cleton-Jansen, A.M., Berx, G., de Leeuw, W.J., ter Haar, N.T., van Roy, F., Cornelisse, C.J., Peterse, J.L., and van de Vijver, M.J. (1997) E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. *Br J Cancer*, **76**(9): 1131–3.

Waldron, D. (2016) Cancer genomics: A multi-layer omics approach to cancer. *Nat Rev Genet*, **17**(8): 436–437.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, **38**(18): e178.

Wang, X. and Simon, R. (2013) Identification of potential synthetic lethal genes to p53 using a computational biology approach. *BMC Medical Genomics*, **6**(1): 30.

Wappett, M. (2014) Bisep: Toolkit to identify candidate synthetic lethality. r package version 2.0.

Wappett, M., Dulak, A., Yang, Z.R., Al-Watban, A., Bradford, J.R., and Dry, J.R. (2016) Multi-omic measurement of mutually exclusive loss-of-function enriches for candidate synthetic lethal gene pairs. *BMC Genomics*, **17**: 65.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., *et al.* (2015) *gplots: Various R Programming Tools for Plotting Data*. R package version 2.17.0.

Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**(6684): 440–2.

Weinstein, I.B. (2000) Disorders in cell circuitry during multistage carcinogenesis: the role of homeostasis. *Carcinogenesis*, **21**(5): 857–864.

Weinstein, J.N., Akbani, R., Broom, B.M., Wang, W., Verhaak, R.G., McConkey, D., Lerner, S., Morgan, M., Creighton, C.J., Smith, C., *et al.* (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, **507**(7492): 315–322.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Chang, K., *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, **45**(10): 1113–1120.

Wickham, H. and Chang, W. (2016) *devtools: Tools to Make Developing R Packages Easier*. R package version 1.12.0.

Wickham, H., Danenberg, P., and Eugster, M. (2017) *roxygen2: In-Line Documentation for R*. R package version 6.0.1.

Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., *et al.* (2004) Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(44): 15682–15687.

World Health Organization (WHO) (2017) Fact sheet: Cancer. http://www.who.int/mediacentre/factsheets/fs297/en/. Updated February 2017, Accessed: 22/03/2017.

Wu, M., Li, X., Zhang, F., Li, X., Kwoh, C.K., and Zheng, J. (2014) In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. *Cancer Inform*, **13**(Suppl 3): 71–80.

Yu, H. (2002) Rmpi: Parallel statistical computing in r. *R News*, **2**(2): 10–14.

Zhang, F., Wu, M., Li, X.J., Li, X.L., Kwoh, C.K., and Zheng, J. (2015) Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *J Bioinform Comput Biol*, **13**(3): 1541002.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., *et al.* (2011) International cancer genome consortium data portala one-stop shop for cancer genomics data. *Database: The Journal of Biological Databases and Curation*, **2011**: bar026.

Zhong, W. and Sternberg, P.W. (2006) Genome-wide prediction of c. elegans genetic interactions. *Science*, **311**(5766): 1481–1484.

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