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

S. Thomas Kelly

31 March 2017

# Methods, Techniques, and Resources

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

## Bioinformatics Resources to Enable Genomics Research

### Public Data and Software Packages

Various bioinformatics resources, such as databases and methods, have become integral parts of genetics and genomics research. Reference genomes, genotyped variants, gene expression, and epigenetics profiles are among the most commonly used resources. Gene expression data in particular is widely available from many microarray and RNA-Seq studies, from repositories such as Gene Expression Omnibus (GEO) , caArray , and ArrayExpress . Such profiles serve as an excellent resource to examine the changes of gene expression occurring in cancers and the variation between samples. These microarray initiatives have set a precedent for data sharing, data mining, and the wider benefits of publicly available data for enabling the scientific community to further utilise the data rather than a single research group or consortium . The practice of integrating findings from publicly available genomics data with the research questions and experimental results of individual research groups has carried over into RNA-Seq datasets including the large-scale cancer genomics projects . This thesis is one such example of an investigation enabled by this wider movement and tools developed in various disciplines to generate, disseminate, and process genomic-scale data.

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

#### Cancer Genome Atlas Data

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

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

Raw count and RSEM normalised TCGA expression data from Illumina RNA-Seq protocols were available from 1,177 samples (113 normal, 1,057 primary tumour, and 7 metastases) for 20,501 genes. TCGA somatic mutation data for 981 samples (976 primary tumours and 5 metastases) across 25,836 genes were available including 969 samples (964 primary tumours and 5 metastases) with corresponding RNA-Seq expression data and 19,166 genes mapped from Ensembl identifiers to gene symbols, of which 16,156 had corresponding gene expression information. Unless otherwise stated, the raw counts were used for further processing rather than the RSEM normalised data (provided by TCGA tier 3).

#### Reactome and Annotation Data

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

## Data Handling

### Normalisation (voom)

Apart from the PAM50 subtyping procedure which required RSEM normalised data (J.S. Parker personal communication), the analysis of the RNA-Seq data presented here was based on raw read count data. Raw read counts were log-scaled, samples were checked for consistency with some removed (as described in section [methods:sample\_qc]) based on a correlation matrix (Euclidean distance), and the final dataset was TMM normalised and then processed using the voom function in the limma R package .

[!ht]

[!ht]

### Sample Triage

The TCGA 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 [fig:corr\_map]. 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 [tab:qc]. These samples showed unusual density plots compared to the rest of the dataset, and exhibited low mean read count in Figures [fig:density] and [fig:boxplot]. A heatmap showing key clinical properties of these excluded samples and their correlation with the remainder of the samples is shown in Figure [fig:corr\_map\_part], and a full correlation heatmap (Figure [fig:corr\_map]) shows these samples as relatively poorly correlated outliers in the bottom rows and left columns. 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.

[!ht]

### Pathway Metagenes and the Singular Value Decomposition

A “metagene” offers a consistent signal of pathway (expression) activation or inactivation by dimension reduction of a matrix, avoiding negatively correlated genes averaging out the signal of a mean-based centroid . Construcing these pathway metagenes used gene sets for Reactome and Gatza signatures (Gatza *et al*., 2011; Gatza *et al*., 2014) as specified above (see Section [methods:gene\_set]). The singular-value decomposition was performed ( where is the data matrix of the gene set with genes samples) and the leading eigenvector (first column of ) corresponding to the largest singular value was used as a metagene for the pathway gene set. To ensure consistent directionality of metagene signals, the median of the gene set in each sample was calculated and correlated against the metagene with the (arbitrary) metagene sign adjusted as needed to conform with the majority of the gene set (i.e., postive correlation between metagene and the median-based centroid). To ensure that genes and pathways were weighted equally, this was performed on a z-transformed dataset of gene expression and samples were scaled (by fractional ranking) for each metagene so that each metagene was on a comparable scale.

#### 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 . The first procedure to compare the SLIPT gene candidates for *CDH1* with an siRNA experimental screen was a direct comparision of the overlapping candidates, presented in a Venn diagram and tested with the test. Since these candidates were not very comparable 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 [methods:permutation]), the same number of genes detected by SLIPT were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1,652 reactome pathways. These over-representation scores () were compared the observed over-representation in the intersection of the SLIPT candidates, with the proportion of resamplings with higher values used for empirical p-values of pathway composition. Pathways for which no resamplings were observed as high as the observed were reported as . These empirical p-values were adjusted for multiple comparisions (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).

## Techniques

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

### Statistical Procedures and Tests

As described in sections [methods:heatmap] and [methods:metagene], the z-transform has been used to generate z-scores in various analyses in this thesis. Here each row of dataset () is transformed into a scores () using the mean () and standard deviation () of the data such that:

. 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 is plotted as or respectively.

Empirical Bayes d ifferential expression analysis was performed using the limma R package . Where specfied, the Fisher’s exact test, test, and correlation were used to measure associations between variables (as implemented in the stats R package ). Unless otherwise specified, Pearson’s correlation was used for correlation analyses () and coefficient of determination (). Where these comparisons are discussed in more detail, Fisher’s exact test and tests are supported by a table or Venn diagram, rendered with the limma R package . In some analyses, correlation is furter supported by a scatter plot and a line of best fit dervied by least squares linear regression.

The t.test function 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 () and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violinplot visualisation (rendered in R).

Multiple comparisons are adjusted by the Benjamini-Hochberg procedure to control the false discovery rate (FDR) unless otherwise specified . This procedure adjusts p-values to achieve an average of the proportion of false-postives among significant tests below a threshold, . The more stringent Holm-Bonferroni (Holm) procedure 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-postive (type-1 error) below a threshold, .

### Gene Set Over-representation Analysis

Gene set enrichment over-representation tests whether there is an enrichment of a gene set (such as 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 [methods:clustering]) or comparison with experimental candidates (see section [methods:venn\_analysis]). Initially, these were performed using the GeneSetDB web tool hosted by the University of Auckland on the Reactome pathways . 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 [methods:venn\_analysis] and [methods:permutation]) performed on version 52 (as described in section [methods:gene\_set]). Thus an implementation of the hypergeometric test in R was used to test for over-representation against Reactome (version 52) pathways. Pathways containing less than 10 genes or more than 500 were excluded before adjusting for multiple comparisons.

### Clustering

Clustering analysis when performed uses unsupervised hierarchical clustering with complete linkage (distance calculated from the furthest possible pairing). For correlation matrices or multivariate normal parameters (e.g., ), the distance metric used was Euclidean distance. For empirical or simulated gene and pathway expression data correlation distance was used, calculated by where is Pearson’s correlation and is the transpose of the expression matrix.

### 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 blue-red scale. Similarly, correlations were plotted on a blue-red scale. These heatmaps were performed using the linkage and distance specified for the clustering performed in Section [methods:clustering]. The gplots R package was used to generate many of the heatmaps throughout this thesis, along with a customised heatmap function (released as heatmap.2x). Where clearly specfied, data have been split into subsets with clustering performed separately on each subset with these plotted alongside each other.

### Modeling and Simulations

Statisical modeling and simulations have been used to test various synthetic lethal detection procedures on simulated data. This involves constructing a statistical model of how synthetic lethality would appear in (continuous normally distributed) gene expression data. Where presented (in section [methods:SL\_Model]), the assumptions of the model are stated clearly. The model allows sampling from a multivariate normal distribution (using the mvtnorm R package ) to generate simulated data with known underlying synthetic lethal partners (detailed in section [methods:simulating SL]). We can test whether statistical procedures, including those developed in this thesis (presented in section [methods:SLIPT]), are capable of detecting them upon this simulated data. This multivariate normal simulation procedure also enables the inclusion of correlation structure which is either given as correlated blocks of genes or derived from pathway structures (as detailed in section [methods:graphsim]).

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

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for , significance and directional criteria met for SLIPT (see section [methods:SLIPT]), and an arbitrary threshold of and for correlation and negative 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 discovery rate” (also important in adjusting for multiple comparisons) was measured here as the proportion of known non-synthetic lethal partners out of all putative partners predicted by the detection procedure. Statistical “accuracy” is the proportion of true predictions for a detection procedure, in this case the total correctly predicted known synthetic lethal partners and correctly negative known non-synthetic lethal partners.

#### Reciever Operating Characteristic (Performance)

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

### Resampling Analysis

Resampling analyses (e.g., “permutation” analysis) are used to statisically test the significance of an observation without assuming the underlying distribution of expected test statisics . 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 candidiates (section [methods:venn\_analysis]) and for assessing the significance of pathway structure among synthetic lethal candidates (section [methods:network\_permutation]).

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

Resampling was performed for comparison (in section [methods:venn\_analysis]) with fixed experimental screen candidates 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 [methods:network\_permutation]) 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 [methods:venn\_analysis]), a million iterations were performed using high performance computing (as detailed in section [methods:HPC]) to provide sufficient accuracy after adjusting for multiple comparisons across pathways. For the purposes of network analysis (in section [methods:network\_permutation]), 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.

## Pathway Structure Methods

### Network and Graph Analysis

Networks are important in considering the structure of relationships in molecular biology, including gene regulation, kinase cellular signaling, and metabolic pathways . 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 . Custom R functions to perform more complex analysis and visualisation of igraph data objects will be described later.

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

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 . Unless otherwise stated, graph structures and networks in thesis will be unweighted and have no parallel edges or loops. Where a directional relationship is known or modelled, it will be represented with a directed edge in a digraph.

### Sourcing Graph Structure Data

Pathway Commons interaction data was sourced using the paxtools-4.3.0 Java application on October 6th 2015 . This utility was used to source ‘sif’ format interaction data into R , from which the human Reactome (version 52) dataset of interactions was imported , 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 bioloigcal pathway.

### 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 [methods:gene\_set]). 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 nodess within the pathway being linked by the surrounding graph structure.

### Network Analysis Metrics

The existing network analysis measures which were applied in this thesis used an implementation in the igraph R package where it was available . Otherwise, custom features were developed for analysis of iGraph objects in R and released as igraph.extensions (as described in section [methods:igraph\_extensions]).

Vertex degree is the number of edges a node has and is a fundamental measure of the importance and connectivity of a network . 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 in the igraph R package . 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 [methods:pathway\_str]).

Network centrality is an alternative measure of the importance or influence of a node to the graph structure . Various strategies are used to derive centrality, typically based on how connected the node is or the impact of node removal on the conenctivity 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 . This is implemented in the igraph R package .

Another network centrality measure that has been previously applied to biological protein interaction networks is the “information centrality”. The information centrality of a node is the relative impact on efficiency (transmission of information via shortest paths) of the network when the node is removed. That is the centrality () for node in graph is defined as:

where is the subgraph with the node removed and is the efficiency derived from shortest paths ( between nodes and ) as:

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.

## Implementation

### Computational Resources and Linux Utilities

Several computers were used to process and store data during this thesis (as summarised in Table [tab:computers]), 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 [tab:computers]) and interacting with these via ZSH shell. However, Ubuntu OS and the Bourne Again SHell (bash) were used at the inception of this project and bash is continues to be used for running scripts. Various Linux applications and command-line utilities were used on these machines (as summarised in Table [tab:computers\_linux]). 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.

### 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 [tab:computers\_r]. 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 developed during this thesis are given in Table [tab:computers\_r\_packages\_dev] with the relevant sections describing their implementation and use where appropriate. Various R packages were used throughout this thesis (as detailed in Table [tab:computers\_r\_packages] with versions specified) these have not been updated if they would change the functionality of the 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 the relevant sections as well. A full table of packages used in this thesis can be found in the Appendix (Table [tab:computers\_r\_packages\_full]).

### High Performance and Parallel Computing

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

The New Zealand eScience Infrastructure (NeSI) is a computating resource providing the Intel Pan cluster hosted by the University of Auckland . 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 [methods:permutation] and [methods:network\_permutation]), calculating information centrality (in section [methods:network\_metrics]), and in simulations (in sections [methods:simulation], [methods:simulation\_SL\_expression], and [methods:graphsim])

Scripts and data were transferred between the Pan cluster and University of Otago computing resources by rsync and the Globus file transfer service . R scripts were run in parallel with the “simple network of workstations” snow R package . This utilised the “message passing interface” when it was feasible with memory requirements to run in parallel across multiple compute nodes, otherwise SOCKS was used to access multiple cores within an instance of R and pass input data to them. R jobs were submitted to queue for avialable resources and run on the Pan cluster via the Slurm workload manager . When running R scripts across many parameters or for memory-intensive jobs, slurm array job submission and independent submission of different parameters via shell commands with arguments passed to R. In some cases, this submission was automated across a range of parameters with Bash scripts.

# Methods Developed During Thesis

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

## A Synthetic Lethal Detection Methodology

[!b]

The SLIPT methodology identifies gene expression patterns consistent with synthetic lethal interactions between a query gene and a panel of candidate interacting partners. Gene expression is called low, medium, or high by separating samples into tertiles (3-quantiles) for each gene. Genes with insufficient expression across all samples were excluded by requiring that the first tertile of raw counts is above zero. Then a test is performed between the query gene and each candidate partner, with the p-values for the test being corrected for multiple testing using false discovery rate (FDR) error control to reduce false positives for large candidate gene panels . Significance was called only if FDR adjusted p-values were below the threshold . A synthetic lethal interaction is predicted (as shown in Figure [fig:SLIPT\_Method]) when (i) the test is significant; (ii) observed low-query, low-candidate samples are less frequent than expected; and (iii) observed low-query, high-candidate and high-query, low-candidate samples are more frequent than expected.

The synthetic lethal prediction procedure has also been adapted to utilise somatic mutation data for the query gene. This is intended to utilise a query gene known to be reccurently mutated in the disease (and dataset), with the majority of mutations inactivating gene function (such as null or framefshift mutations). A synthetic lethal interaction is predicted (as shown in Figure [fig:SLIPT\_Method\_mtSL]) when (i) the test is significant; (ii) observed mutant-query, low-candidate samples are less frequent than expected; and (iii) observed mutant-query, high-candidate and wild-type-query, low-candidate samples are more frequent than expected. Unless otherwise specified, computationally predicted synthetic lethal gene candidiates from SLIPT used expression data (exprSL) for both genes (as shown in Figure [fig:SLIPT\_Method]) rather than mutation data (mtSL) for the query gene (as shown in Figure [fig:SLIPT\_Method\_mtSL]).

[!ht]

## Simulations and Modelling of Synthetic Lethality in Expression Data

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

### A Model of Synthetic Lethality

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

[!ht]

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

[!ht]

The first condition of the synthetic lethal model is that if there are only two synthetic lethal genes (e.g., *CDH1* and one SL partner), then they will not both be non-functional in the same sample (in an ideal model). Gene function is thus determined for each sample in a model of synthetic lethal with the proportion of samples with a functional or non-functional gene being arbitrary. Whether a gene is functional can similarly be modelled by an arbitrary threshold of continuous and normally distributed gene expression data to define gene function (as shown in Figure [fig:SL\_Model\_Expression]). For the purposes of modeling synthetic lethality in breast cancer expression data, a threshold of the 30th percentile of the expression levels was used because approximately 30% of samples analysed had *CDH1* inactivation. This was generalised for a model of the proportion of samples inactivated for each gene. In this ideal case, we would not expect to observe any samples lowly expressing both of these genes. While this is not observed, that is to be expected as it is unlikely that only 2 genes will have an exclusive synthetic lethal partnership. The threshold of the 0.3 quantile was used in simulations dervied from this model throughout this thesis.

[!ht]

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 [fig:SL\_Model\_Higher]. Even when testing pairwise interactions, modelling higher level interactions that may interfere is important. If there are additional synthetic lethal partners, there are two possibilities for adding these: 1) that they are independent partners of the query genes interacting pairwise (and not with each other) or 2) that an addition partner gene interacts with both of the synthetic lethal genes already in the system and any of the 3 (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 thus we make the more stringent assumption that all synthetic lethal partner genes interact with each other: that only one of these must be expressed to satisfy the model of synthetic lethality. In this model any of the synthetic lethal genes in a higher-order interaction is able to provide the missing function of the others, allowing for higher-level synthetic lethal partners to compensate for loss a synthetic lethal gene pair. While samples expressing low levels of the synthetic lethal gene pairs will be under-represented, they may not be completely absent from the dataset due to these higher-level interactions.

In the example of 3 synthetic lethal genes [fig:SL\_Model\_Higher], only one of genes involved in the higher-order synthetic lethal interaction is required for cell viability. Thus, if we consider synthetic lethal pairs, only a subset of these samples will be inviable (i.e., removed from simulated data), leading to an under-representation.

In practice, samples are not removed from a simulated dataset, rather the expression and function of the query gene is generated across samples separately from the pool of potential partner genes. The query gene data is matched to simulated samples (as shown in Figure [fig:simulate\_add\_query]), satisfying the synthetic lethal condition with the procedure described in section [methods:simulating SL]. This is performed to maintain a comparable samples size across simulations and the preserve the assumed (multivariate) normal distribution of the data.

### Simulation Procedure

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

This also has the added advantage of being amenable to simulating correlation structures with the multivariate normal distribution (using the mvtnorm R package ). The parameter is a covariance matrix defines the correlation structure between simulated genes being sampled. With the diagonal of the matrix is one, this simulates genes with a standard deviation of one and the covariance parameters between them are the correlations between each gene. In Figure [fig:simulate\_function], an example of such a simulated multivariate normal dataset is shown with the functional threshold applied.

[!ht]

Once we have generated a simulated dataset, the samples are compared by gene function (as derived from a functional threshold). Known underlying synthetic lethal partners are selected within the dataset and a query gene is generated by sampling from the normal distribution. These are matched (as shown for 2 synthetic lethal partners in Figure [fig:simulate\_add\_query]) such that the synthetic lethal condition is met: that at least one of the synthetic partner genes and the query gene are functional in any particular cell. This is done by ordering the samples 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.

[!ht]

This results a simulated dataset where samples with non-functional query gene do not have loss of function in all of the synthetic lethal partners. At least one partner gene was required to be functional in each sample. Similarly, the query gene is functional in all samples where all of the synthetic lethal partner genes are permitted to be non-functional. Therefore we have generated a dataset with known synthetic lethal partners (see Figure [fig:simulate\_SL]) by as few assumptions about the relationships between the each synthetic lethal pair as possible (and allowing compensating functions from higher-order interactions). This has been designed to have the most stringent (least detectable) synthetic lethal relationships where higher-order interactions are possible for the purposes of testing pairwise detection procedures such as SLIPT.

[!ht]

image

## Detecting Simulated Synthetic Lethal Partners

Test

### Multivariate Normal Simulation of Synthetic lethality

(using the mvtnorm R package )

#### Simulation with Directional Testing

Test

#### Simulation with Query-Correlated Pathways

Test

#### Simulated Expression Heatmaps

Test

### Replication Simulation Heatmap

Test

## Graph Structure Methods

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

### Upstream and Downstream Gene Detection

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

Using the Reactome version 52 data as described in section [methods:graph\_data], genes identified by each synthetic lethal discovery approach were mapped to the graph structure for the candidate pathways identified in Chapter 4 (with subgraphs defined as described in section [methods:subgraphs]). To test whether siRNA candidate genes were upstream of SLIPT candidate genes, shortest paths were traced between each potential pair of these genes in a directed network. The number of genes where the siRNA candidate was upstream were scored “up” and where the siRNA candidate was downstream were scored “down”. This procedure enabled counting the total number of shortest paths which supported siRNA genes being upstream or downstream of the SLIPT genes and measuring the difference between these to determine if there is an imblance in a particular direction. While indicative of the number of paths between the gene candidate groups in either direction, this is alone is not indicative of a statistically significant stucture or relationships between siRNA and SLIPT genes, however it may be combined with a permutation resampling procedure (as described in section [methods:network\_permutation]) to test for directional relationships in either direction.

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

#### 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 permutated genes are measured for pathway structure between the permutated groups as performed for the observed candidiates (as performed in section [methods:pathway\_str]). Permuting iteratively over these pathways generates a distribution of pathway structure relationships expected by chance. This null distribution can be compared to the observed counts of relationships (in either direction) which yields a permutation p-value as the proportion of permutations in which had value or greater or more extreme magnitude than the observed value.

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

This permutation procedure does not assume the underlying distribution of the data under the null hypothesis and accounts for the total number of nodes, edges, siRNA, and SLIPT genes in each pathway network structure. The intersection size of the siRNA and SLIPT genes was originally not accounted for under the shortest path counts procedure that excluded them. A refined version of this procedure ensured that the number of intersecting genes was match the number observed to test for pathway structure without changing the intersection size, the subject of prior analyses.

#### Ranking Based on Biological Context

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

A network-based approach was used to determine the pathway ranking of genes in a computationally rational way when applied to different biological pathways with a directed graph structure, (without loops). The diameter of the network (length of longest possible shortest path between the most distant genes) was determined and a gene at the downstream end of the pathway is selected as a gene () at the end of a diameter spanning shortest path, assigned a rank of:

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

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

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

This ranking of pathway directionality can be used for comparison with measures of the number of genes of each candidate group and the support for being synthetic lethal partners with either approach.

### Simulating Gene Expression from Graph Structures

A further refinement of the simulution procedure is to generate expression data with correlation structure derived from a known graph structure. This enables modelling of synthetic lethal partners within a biological pathway and the impact of pathway structure on synthetic lethal prediction to be considered. First a simulated pathway is constructed using a graph structure, with the igraph R package with the added provision of including the state of the edges, that is whether they activate or inhibit downstream pathway members. Here we consider purely whether biological pathway members would be expected to have correlated gene expression (higher than the background of genes in other pathways) but this framework is also applicable to modelling protein levels in a kinase regulation cascade or metabolic pathway with related substrates and products.

[!hp]

image

First we must consider the graph structure upon which simulated data will be generated (by sampling from a multivariate normal distribution using the mvtnorm R package ). Throughout this section, the simulation procedure will be demonstrated with the relatively simple constructed graph structure shown in Figure [fig:simple\_graph]. This graph structure visualisation was developed specifically for (directed) igraph objects in R and has been released in the plot.igraph package and igraph.extensions library (see Table [tab:computers\_r\_packages\_dev] and section [methods:igraph\_extensions]). The plot\_directed function allows customisation of plot parameters for each node or edge and mixed (directed) edge types for indicating activation or inhibition. These inhibition links (which often occur in biological pathways) are demonstrated in Figure [fig:simple\_graph:second].

[!hp]

image

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

A matrix is derived from this distance weighting matrix by creating a matrix (with a diagonal of ) where each node has a variance and standard deviation of 1 such that covariances between adjacent nodes are assigned by a correlation parameter (cor), and the remaining matrix based on weighting these correlations with by the distance matrix (or the nearest positive definite matrix). For the purposes of this thesis the correlation parameter, , unless otherwise specified as used for the example in Figure [fig:simulation\_activating:third]. This matrix is used to sample from a multivariate normal distribution where each node has a mean of , standard deviation and covariance within the range such that they are correlations. This generates a simulated (continuous normally distributed) expression profile for each node (as shown in Figure [fig:simulation\_activating:fourth]) with corresponding correlation structure (Figure [fig:simulation\_activating:fifth]). This simulated correlation structure closely resembles the expected correlation structure (Sigma in [fig:simulation\_activating:third]) even for the relatively modest sample size () illustrated in [fig:simulation\_activating]. Once a simulated gene expression dataset has been generated (as in Figure [fig:simulation\_activating:fourth]) then a discrete matrix of gene function can be constructed with a functional threshold quantile (the parameter pr) to simulate functional relationships of synthetic lethality. For the purposes of this thesis the parameter pr is set to the 0.3 quantile which generates functional discrete matrices such as those used for synthetic lethal simulation in section [methods:simulating SL] (as shown Figure [fig:simulation\_activating:sixth])

[!hp]

image

The simulation procedure discussed in Figure [fig:simulation\_activating] is amenable to pathways containing inhibition links (as shown in Figure [fig:simulation\_inhibiting]) with a few refinements. With the inhibition links (as shown in Figure [fig:simulation\_inhibiting:first]), the distances are calculated in the same manner as before (Figure [fig:simulation\_inhibiting:second]) but the inhibitions are accounted for by iteratively multiplying downstream nodes by to form blocks of negative correlations (as shown in Figures [fig:simulation\_inhibiting:third] and [fig:simulation\_inhibiting:fifth]). As before, a multivariate normal distribution with these negative correlations can be sampled to generate simulated data (as shown in Figures [fig:simulation\_inhibiting:fourth] and [fig:simulation\_inhibiting:sixth]).

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

## Customised Functions and Packages Developed

[Move to Appendix?]

Various R packages have been developed throughout this thesis using devtools and roxygen to enable reproducibility of customised analysis and visualisation. Many of these have the added benefit of the functions being documented, demonstrated in example Vignettes, and released on GitHub to enable the research community to access utilise them in their own analysis. These are summarised in Table [tab:computers\_r\_packages\_dev] along with the corresponding urls for their GitHub repository which contains a README with instructions for installation with the devtools R package and links to the Vignette(s) where available.

### Synthetic Lethal Interaction Prediction Tool (SLIPT)

The statistical methodology for detection of synthetic lethality in gene expression data is one of the main procedures developed in this thesis, as described in section [methods:SLIPT]. 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 .

slipt is amenable to analysis of any effectively continuous measure of gene activity (e.g., microarray, RNA-Seq, protein abundance, or pathway metagenes). Executing it is simple, the prep\_data\_for\_SL function performs binning of samples into “low”, “medium”, and “high” categories for each gene and the detect\_SL function tests all potential partners of a given gene by performing the chi-squared test and directional conditions which returns a table summarising the observed and expected sample numbers used for these criteria, the values, and corresponding p-values including adjusting for multiple comparisons. The count\_of\_SL and table\_of\_SL serve to facilitate summary and extracting the positive SLIPT hits respectively from the table of predictions of synthetic lethal partners.

The SLIPT methodology in this package release has been used in later analyses rather than the corresponding source R code, including it’s use on remote machines and upon simulated data. In particular the functions in the package facilitate alterations to parameters such as the proportion of samples called as exhibiting low or high gene activity. This release will also enable use of SLIPT in future investigations for reproducibility and applications to other genes or diseases.

### Data Visualisation

Customisations to existing data visualisations in R have been developed to present data throughout this thesis. vioplotx and heatmap.2x are refined versions of the vioplot package and heatmap.2 provided by the gplots package .

[!p]

image

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

[!p]

The heatmap.2x provides extensions for annotation colour bars for both the rows and columns (as shown in Figure [fig:heatmap.2x]). Multiple bars are enabled on both axes with matrix inputs (rather than single vector for heatmap.2) which facilitates additional plotting of gene and sample characteristics for comparison with correlation matrices, expression profiles, or pathway metagenes. Annotation bar inputs correspond to their orientation on the plot with each colour bar being provided as a column for the row annotation on the left of the heatmap and each colour bar provided 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 suppression of reordering columns for if samples are ordered in advance (e.g., ranked by a metagene or split into groups clustered separately). These features are used through this thesis and are provided in a package [GitHub repository](https://github.com/TomKellyGenetics/heatmap.2x).

### Extensions to the iGraph Package

The following additional features were developed during this thesis using “iGraph” data objects, building upon the igraph package . These have been released as separate packages for each respective procedure or installed together as a collection of extensions to the igraph package.

#### Sampling Simulated Data from Graph Structures

The graphsim package implements the procedure for simulating gene expression from graph structures (as described in section [methods:graphsim]). 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 supported as well as deriving correlation structures from adjacency matrices or weighted by the number of common links (i.e., size of the shared “neighbourhood” ) between each pair of nodes. Functions to compute these are called directly by passing paramters to them when running generate\_expression or the relevant make\_sigma\_mat. This enables simulating expression data directly from a graph structure, with the intermediate steps automated and generating parameters for mvtnorm from graph structures or matrices derived from them. These functions support the use of a “state” parameter for assigning activating or inhibiting to each edge.

#### Plotting Directed Graph Structures

The plot.igraph package provides the plot\_directed function specifically developed to handle directed graph structures and plot activating or inhibiting for each edge (as described in section [methods:graphsim]). As shown in Figure [fig:simple\_graph2], this function supports customisation of 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 textsize of labels. The “state” parameter for assigning activating or inhibiting to each edge determines whether edges are depected with 30 or 90 arrowheads but colours are assigned separately and may be customised. Customised parameters across nodes or edges are given as vectors but scalar parameters can also be used to apply the same ploting parameters across them. The default layout function is layout.fruchterman.reingold but any layout function supported by plot function in igraph is compatible such as the layout.kamada.kawai used to implement the KamadaKawai algorithm for graph plots throughout this thesis.

[!hp]

image

#### Computing Information Centrality

As described in section [methods:network\_metrics], the shortest paths of a network are computed by the igraph package which can be extended to calculate the network efficiency but is not provided by the package itself. The “information centrality” of a vertex is computed as the relative change in the network efficiency when the vertex is removed. This is performed iteratively for each node and the sum of information centrality for each vertex is the information centrality for the network. These metrics are provided by the info.centrality package.

#### Testing Pathway Structure with Permutation Testing

Computing the number of shortest paths in either direction between two different groups of nodes within a graph provides a means to test pathway structure relationships, such as a comparison of siRNA and SLIPT candidate genes in a pathway structure. This pathway relationship metric was implemented in the pathway.stucture.permutation package with permutation testing (as described in sections [methods:pathway\_str] and [methods:network\_permutation]).

#### Metapackage to Install iGraph Functions

These features may be installed together with the igraph.extensions meta package which can be accessed from a [GitHub repository](https://github.com/TomKellyGenetics/igraph.extensions). This installs igraph and the packages described in section [methods:igraph\_extensions] including their dependencies for matrix operations and statistical procedures: Matrix, matrixcalc, and mvtnorm .

# Sample Correlation

[!ht]

[!ht]

# Software Used for Thesis

|llllll|

& & & & &  
base & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
abind & CRAN & & 1.4-5 & & 1.4-3  
acepack & CRAN & & 1.4.1 & & 1.3-3.3  
ade4 & CRAN & & 1.7-5 & &  
annaffy & Bioconductor & & 1.46.0 & &  
AnnotationDbi & Bioconductor & & 1.36.0 & 1.36.0 & 1.34.4  
apComplex & CRAN & & 2.40.0 & &  
ape & CRAN & & 4 & & 3.4  
arm & CRAN & & 1.9-3 & &  
assertthat & CRAN & 0.1 & 0.1 & 0.1 & 0.1  
backports & CRAN & 1.0.5 & 1.0.4 & 1.0.5 & 1.0.2  
base64 & CRAN & & & 2 & 2  
base64enc & CRAN & & 0.1-3 & & 0.1-3  
beanplot & CRAN & & 1.2 & 1.2 & 1.2  
BH & CRAN & 1.60.0-2 & 1.62.0-1 & 1.62.0-1 & 1.60.0-2  
Biobase & Bioconductor & & 2.34.0 & 2.34.0 & 2.32.0  
BiocGenerics & Bioconductor & & 0.20.0 & 0.20.0 & 0.18.0  
BiocInstaller & Bioconductor & & 1.24.0 & 1.20.3 & 1.22.3  
BiocParallel & Bioconductor & & 1.8.1 & 1.8.1 &  
Biostrings & Bioconductor & & 2.42.1 & 2.42.0 &  
BiSEp & Bioconductor & & 2.0.1 & 2.0.1 & 2.0.1  
bitops & CRAN & 1.0-6 & 1.0-6 & 1.0-6 & 1.0-6  
boot & base & 1.3-18 & 1.3-18 & 1.3-18 & 1.3-18  
brew & CRAN & 1.0-6 & 1.0-6 & 1.0-6 & 1.0-6  
broom & CRAN & 0.4.1 & & &  
caTools & CRAN & 1.17.1 & 1.17.1 & 1.17.1 & 1.17.1  
cgdsr & CRAN & & 1.2.5 & &  
checkmate & CRAN & & 1.8.2 & & 1.7.4  
chron & CRAN & 2.3-47 & 2.3-48 & 2.3-50 & 2.3-47  
class & base & 7.3-14 & 7.3-14 & 7.3-14 & 7.3-14  
cluster & base & 2.0.5 & 2.0.5 & 2.0.5 & 2.0.4  
coda & CRAN & & 0.19-1 & & 0.18-1  
codetools & base & 0.2-15 & 0.2-15 & 0.2-15 & 0.2-14  
colorRamps & CRAN & & 2.3 & &  
colorspace & CRAN & 1.2-6 & 1.3-2 & 1.3-2 & 1.2-6  
commonmark & CRAN & 1.1 & & 1.2 &  
compiler & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
corpcor & CRAN & & 1.6.8 & 1.6.8 & 1.6.8  
Cprob & CRAN & & 1.2.4 & &  
crayon & CRAN & 1.3.2 & 1.3.2 & 1.3.2 & 1.3.2  
crop & CRAN & & 0.0-2 & 0.0-2 &  
curl & CRAN & 1.2 & 2.3 & 2.3 & 0.9.7  
d3Network & CRAN & & 0.5.2.1 & &  
data.table & CRAN & 1.9.6 & 1.10.0 & 1.10.1 & 1.9.6  
data.tree & CRAN & & 0.7.0 & 0.7.0 &  
datasets & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
DBI & CRAN & 0.5-1 & 0.5-1 & 0.5-1 & 0.5-1  
dendextend & CRAN & 1.4.0 & 1.4.0 & 1.4.0 &  
DEoptimR & CRAN & 1.0-8 & 1.0-8 & 1.0-8 & 1.0-4  
desc & CRAN & 1.1.0 & & 1.1.0 &  
devtools & CRAN & 1.12.0 & 1.12.0 & 1.12.0 & 1.12.0  
DiagrammeR & CRAN & & 0.9.0 & 0.9.0 &  
dichromat & CRAN & 2.0-0 & 2.0-0 & 2.0-0 & 2.0-0  
digest & CRAN & 0.6.10 & 0.6.11 & 0.6.12 & 0.6.9  
diptest & CRAN & 0.75-7 & 0.75-7 & 0.75-7 &  
doParallel & CRAN & 1.0.10 & 1.0.10 & 1.0.10 & 1.0.10  
dplyr & CRAN & 0.5.0 & 0.5.0 & 0.5.0 & 0.5.0  
ellipse & CRAN & & 0.3-8 & 0.3-8 & 0.3-8  
evaluate & CRAN & & 0.1 & 0.1 & 0.9  
fdrtool & CRAN & & 1.2.15 & &  
fields & CRAN & & 8.1 & &  
flexmix & CRAN & 2.3-13 & 2.3-13 & 2.3-13 &  
forcats & CRAN & 0.2.0 & & &  
foreach & CRAN & 1.4.3 & 1.4.3 & 1.4.3 & 1.4.3  
foreign & base & 0.8-67 & 0.8-67 & 0.8-67 & 0.8-66  
formatR & CRAN & & 1.4 & 1.4 & 1.4  
Formula & CRAN & & 1.2-1 & & 1.2-1  
fpc & CRAN & 2.1-10 & 2.1-10 & 2.1-10 &  
futile.logger & CRAN & & 1.4.3 & 1.4.3 & 1.4.1  
futile.options & CRAN & & 1.0.0 & 1.0.0 & 1.0.0  
gdata & CRAN & 2.17.0 & 2.17.0 & 2.17.0 & 2.17.0  
geepack & CRAN & & 1.2-1 & &  
GenomeInfoDb & Bioconductor & & 1.10.2 & 1.10.1 &  
GenomicAlignments & Bioconductor & & 1.10.0 & 1.10.0 &  
GenomicRanges & Bioconductor & & 1.26.2 & 1.26.1 &  
ggm & CRAN & & 2.3 & &  
ggplot2 & CRAN & 2.1.0 & 2.2.1 & 2.2.1 & 2.1.0  
git2r & CRAN & 0.15.0 & 0.18.0 & 0.16.0 & 0.15.0  
glasso & CRAN & & 1.8 & &  
GO.db & Bioconductor & & 3.4.0 & 3.2.2 & 3.3.0  
GOSemSim & Bioconductor & & 2.0.3 & 1.28.2 & 1.30.3  
gplots & CRAN & 3.0.1 & 3.0.1 & 3.0.1 & 3.0.1  
graph & Bioconductor & & 1.52.0 & &  
graphics & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
graphsim &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.1.0 & 0.1.0 & 0.1.0 & 0.1.0  
grDevices & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
grid & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
gridBase & CRAN & 0.4-7 & 0.4-7 & 0.4-7 & 0.4-7  
gridExtra & CRAN & 2.2.1 & 2.2.1 & 2.2.1 & 2.2.1  
gridGraphics & CRAN & & 0.1-5 & &  
gtable & CRAN & 0.2.0 & 0.2.0 & 0.2.0 & 0.2.0  
gtools & CRAN & 3.5.0 & 3.5.0 & 3.5.0 & 3.5.0  
haven & CRAN & 1.0.0 & & &  
heatmap.2x &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.0.0.9000 & 0.0.0.9000 & 0.0.0.9000 & 0.0.0.9000  
hgu133plus2.db & Bioconductor & & 3.2.3 & &  
highr & CRAN & & 0.6 & 0.6 & 0.6  
Hmisc & CRAN & & 4.0-2 & 4.0-2 & 3.17-4  
hms & CRAN & 0.2 & 0.3 & &  
htmlTable & CRAN & & 1.8 & 1.9 &  
htmltools & CRAN & 0.3.5 & 0.3.5 & 0.3.5 & 0.3.5  
htmlwidgets & CRAN & & 0.8 & 0.8 &  
httpuv & CRAN & 1.3.3 & & 1.3.3 &  
httr & CRAN & 1.2.1 & 1.2.1 & 1.2.1 & 1.1.0  
huge & CRAN & & 1.2.7 & &  
hunspell & CRAN & & 2.3 & & 2  
hypergraph & CRAN & & 1.46.0 & &  
igraph & CRAN & 1.0.1 & 1.0.1 & 1.0.1 & 1.0.1  
igraph.extensions &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.1.0.9001 & 0.1.0.9001 & 0.1.0.9001 & 0.1.0.9001  
influenceR & CRAN & & 0.1.0 & 0.1.0 &  
info.centrality &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.1.0 & 0.1.0 & 0.1.0 & 0.1.0  
IRanges & Bioconductor & & 2.8.1 & 2.8.1 & 2.6.1  
irlba & CRAN & 2.1.1 & 2.1.2 & 2.1.2 & 2.0.0  
iterators & CRAN & 1.0.8 & 1.0.8 & 1.0.8 & 1.0.8  
jpeg & CRAN & & 0.1-8 & &  
jsonlite & CRAN & 1.1 & 1.2 & 1.3 & 0.9.20  
KEGG.db & Bioconductor & & 3.2.3 & &  
kernlab & CRAN & 0.9-25 & 0.9-25 & 0.9-25 &  
KernSmooth & base & 2.23-15 & 2.23-15 & 2.23-15 & 2.23-15  
knitr & CRAN & & 1.15.1 & 1.15.1 & 1.14  
labeling & CRAN & 0.3 & 0.3 & 0.3 & 0.3  
lambda.r & CRAN & & 1.1.9 & 1.1.9 & 1.1.7  
lattice & base & 0.20-34 & 0.20-34 & 0.20-34 & 0.20-33  
latticeExtra & CRAN & & 0.6-28 & & 0.6-28  
lava & CRAN & & 1.4.6 & &  
lavaan & CRAN & & 0.5-22 & &  
lazyeval & CRAN & 0.2.0 & 0.2.0 & 0.2.0 & 0.2.0  
les & CRAN & & 1.24.0 & &  
lgtdl & CRAN & & 1.1.3 & &  
limma & Bioconductor & & 3.30.7 & 3.30.3 &  
lme4 & CRAN & & 1.1-12 & & 1.1-12  
lubridate & CRAN & 1.6.0 & & &  
magrittr & CRAN & 1.5 & 1.5 & 1.5 & 1.5  
maps & CRAN & & 3.1.1 & &  
markdown & CRAN & & 0.7.7 & 0.7.7 & 0.7.7  
MASS & base & 7.3-45 & 7.3-45 & 7.3-45 & 7.3-45  
Matrix & base & 1.2-7.1 & 1.2-7.1 & 1.2-8 & 1.2-6  
matrixcalc & CRAN & 1.0-3 & 1.0-3 & 1.0-3 & 1.0-3  
mclust & CRAN & 5.2 & 5.2.1 & 5.2.2 & 5.2  
memoise & CRAN & 1.0.0 & 1.0.0 & 1.0.0 & 1.0.0  
methods & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
mgcv & base & 1.8-16 & 1.8-16 & 1.8-17 & 1.8-12  
mi & CRAN & & 1 & &  
mime & CRAN & 0.5 & 0.5 & 0.5 & 0.4  
minqa & CRAN & & 1.2.4 & & 1.2.4  
mnormt & CRAN & 1.5-5 & 1.5-5 & & 1.5-4  
modelr & CRAN & 0.1.0 & & &  
modeltools & CRAN & 0.2-21 & 0.2-21 & 0.2-21 &  
multtest & Bioconductor & & 2.30.0 & 2.30.0 &  
munsell & CRAN & 0.4.3 & 0.4.3 & 0.4.3 & 0.4.3  
mvtnorm & CRAN & 1.0-5 & 1.0-5 & 1.0-6 & 1.0-5  
network & CRAN & & 1.13.0 & &  
nlme & base & 3.1-128 & 3.1-128 & 3.1-131 & 3.1-128  
nloptr & CRAN & & 1.0.4 & & 1.0.4  
NMF & CRAN & 0.20.6 & 0.20.6 & 0.20.6 & 0.20.6  
nnet & base & 7.3-12 & 7.3-12 & 7.3-12 & 7.3-12  
numDeriv & CRAN & & 2016.8-1 & & 2014.2-1  
openssl & CRAN & 0.9.4 & 0.9.6 & 0.9.6 & 0.9.4  
org.Hs.eg.db & Bioconductor & & 3.1.2 & & 3.3.0  
org.Sc.sgd.db & Bioconductor & & 3.4.0 & &  
parallel & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0

R Packages used during Thesis

|  |
| --- |
| pathway.structure |
| .permutation |

&

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.1.0 & 0.1.0 & 0.1.0 & 0.1.0  
pbivnorm & CRAN & & 0.6.0 & &  
PGSEA & Bioconductor & & 1.48.0 & &  
pkgmaker & CRAN & 0.22 & 0.22 & 0.22 & 0.22  
PKI & CRAN & & 0.1-3 & &  
plogr & CRAN & & 0.1-1 & 0.1-1 &  
plot.igraph &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.0.0.9001 & 0.0.0.9001 & 0.0.0.9001 & 0.0.0.9001  
plotrix & CRAN & & 3.6-4 & &  
plyr & CRAN & 1.8.4 & 1.8.4 & 1.8.4 & 1.8.3  
png & CRAN & & 0.1-7 & & 0.1-7  
prabclus & CRAN & 2.2-6 & 2.2-6 & 2.2-6 &  
praise & CRAN & 1.0.0 & 1.0.0 & & 1.0.0  
pROC & CRAN & & 1.8 & 1.9.1 &  
prodlim & CRAN & & 1.5.7 & &  
prof.tree & CRAN & & 0.1.0 & &  
proftools & CRAN & & 0.99-2 & &  
progress & CRAN & & & 1.1.2 &  
psych & CRAN & 1.6.12 & 1.6.12 & &  
purrr & CRAN & 0.2.2 & 0.2.2 & 0.2.2 & 0.2.2  
qgraph & CRAN & & 1.4.1 & &  
quadprog & CRAN & & 1.5-5 & 1.5-5 & 1.5-5  
R.methodsS3 & CRAN & & 1.7.1 & & 1.7.1  
R.oo & CRAN & & 1.21.0 & & 1.20.0  
R.utils & CRAN & & 2.5.0 & &  
R6 & CRAN & 2.1.3 & 2.2.0 & 2.2.0 & 2.1.3  
RBGL & CRAN & & 1.50.0 & &  
RColorBrewer & CRAN & 1.1-2 & 1.1-2 & 1.1-2 & 1.1-2  
Rcpp & CRAN & 0.12.7 & 0.12.9 & 0.12.9 & 0.12.7  
RcppArmadillo & CRAN & & & 0.7.700.0.0 & 0.6.700.6.0  
RcppEigen & CRAN & & 0.3.2.9.0 & & 0.3.2.8.1  
RCurl & CRAN & & 1.95-4.8 & 1.95-4.8 & 1.95-4.8  
reactome.db & Bioconductor & & 1.52.1 & 1.52.1 &  
reactometree &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

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

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.1.0 & 0.1.0 & 0.1.0 & 0.1.0  
sm & CRAN & 2.2-5.4 & 2.2-5.4 & &  
sna & CRAN & & 2.4 & &  
snow & CRAN & 0.4-1 & 0.4-2 & 0.4-2 & 0.3-13  
sourcetools & CRAN & 0.1.5 & & 0.1.5 &  
SparseM & CRAN & & 1.74 & & 1.7  
spatial & base & 7.3-11 & 7.3-11 & 7.3-11 & 7.3-11  
splines & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
statnet.common & CRAN & & 3.3.0 & &  
stats & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
stats4 & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
stringi & CRAN & 1.1.1 & 1.1.2 & 1.1.2 & 1.0-1  
stringr & CRAN & 1.1.0 & 1.1.0 & 1.2.0 & 1.0.0

R Packages used during Thesis

|  |
| --- |
| Summarized |
| Experiment |

& Bioconductor & & 1.4.0 & 1.4.0 &  
survival & base & 2.39-4 & 2.40-1 & 2.40-1 & 2.39-4  
tcltk & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
testthat & CRAN & 1.0.2 & 1.0.2 & & 1.0.2  
tibble & CRAN & 1.2 & 1.2 & 1.2 & 1.2  
tidyr & CRAN & 0.6.1 & 0.6.1 & 0.6.1 &  
tidyverse &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| hadley |

& 1.1.1 & & &  
timeline & CRAN & & 0.9 & &  
tools & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
tpr & CRAN & & 0.3-1 & &  
trimcluster & CRAN & 0.1-2 & 0.1-2 & 0.1-2 &  
Unicode & CRAN & 9.0.0-1 & 9.0.0-1 & 9.0.0-1 &  
utils & base & 3.3.2 & 3.3.2 & 3.3.1 & 3.3.0  
vioplot & CRAN & & 0.2 & &  
vioplotx &

R Packages used during Thesis

|  |
| --- |
| GitHub |
| TomKellyGenetics |

& 0.0.0.9000 & 0.0.0.9000 & &  
viridis & CRAN & 0.3.4 & 0.3.4 & 0.3.4 &  
visNetwork & CRAN & & 1.0.3 & 1.0.3 &  
whisker & CRAN & 0.3-2 & 0.3-2 & 0.3-2 & 0.3-2  
withr & CRAN & 1.0.2 & 1.0.2 & 1.0.2 & 1.0.2  
XML & base & 3.98-1.3 & 3.98-1.1 & 3.98-1.5 & 3.98-1.4  
xml2 & CRAN & 1.1.1 & & 1.1.1 & 1.0.0  
xtable & CRAN & 1.8-2 & 1.8-2 & 1.8-2 & 1.8-2  
XVector & Bioconductor & & 0.14.0 & 0.14.0 &  
yaml & CRAN & & 2.1.14 & 2.1.14 & 2.1.13  
zlibbioc & CRAN & & 1.20.0 & 1.20.0 &  
zoo & CRAN & 1.7-13 & 1.7-14 & & 1.7-13

# Source Code for thesis.dvi

## thesis.tex

### abstract.tex

### acknowledgements.tex

### intro.tex

### literature.tex

### new\_ideas.tex

### implementation.tex

### results.tex

### conclusion.tex

### appendices.tex

# Secondary Screen Data

A series of experimental genome-wide siRNA screens have been performed on synthetic lethal partners of *CDH1* . The strongest candidates from a primary screen were subject to a further secondary screen for validation by independent replication with 4 gene knockdowns with different targeting siRNA. As shown in Table [tab:secondary\_screen], there is significant ( by Fisher’s exact test) association between SLIPT candidates and stronger validations of siRNA candidates. Since there were more SLIPT genes among those not validated and more SLIPT genes among those validated with several siRNAs, this supports the use of SLIPT as a synthetic lethal discovery procedure which may augment such screening experiments.

[!ht]

rrcccccl & & &  
& & 0/4 & 1/4 & 2/4 & 3/4 & & Total  
& & 70 & 46 & 31 & 8 & &  
& & 85 & 44 & 10 & 4 & &  
& & 190 & 90 & 31 & 10 & &  
& & 175 & 91 & 42 & 12 & &  
& & & & & & &