2017-05-12 03:31:00

## Contents

[Glossary](#Q1-1-4)  
[Acronyms](#Q1-1-6)  
2 [Methods and Resources](#x1-60002)  
 2.1 [Bioinformatics Resources for Genomics Research](#x1-70001)  
  2.1.1 [Public Data and Software Packages](#x1-80001)  
   2.1.1.1 [Cancer Genome Atlas Data](#x1-90001)  
   2.1.1.2 [Reactome and Annotation Data](#x1-100002)  
 2.2 [Data Handling](#x1-110002)  
  2.2.1 [Normalisation](#x1-120001)  
  2.2.2 [Sample Triage](#x1-130002)  
  2.2.3 [Metagenes and the Singular Value Decomposition](#x1-140003)  
   2.2.3.1 [Candidate Triage and Integration with Screen Data](#x1-150001)  
 2.3 [Techniques](#x1-160003)  
  2.3.1 [Statistical Procedures and Tests](#x1-170001)  
  2.3.2 [Gene Set Over-representation Analysis](#x1-180002)  
  2.3.3 [Clustering](#x1-190003)  
  2.3.4 [Heatmap](#x1-200004)  
  2.3.5 [Modeling and Simulations](#x1-210005)  
   2.3.5.1 [Receiver Operating Characteristic (Performance)](#x1-220001)  
  2.3.6 [Resampling Analysis](#x1-230006)  
 2.4 [Pathway Structure Methods](#x1-240004)  
  2.4.1 [Network and Graph Analysis](#x1-250001)  
  2.4.2 [Sourcing Graph Structure Data](#x1-260002)  
  2.4.3 [Constructing Pathway Subgraphs](#x1-270003)  
  2.4.4 [Network Analysis Metrics](#x1-280004)  
 2.5 [Implementation](#x1-290005)  
  2.5.1 [Computational Resources and Linux Utilities](#x1-300001)  
  2.5.2 [R Language and Packages](#x1-310002)  
  2.5.3 [High Performance and Parallel Computing](#x1-320003)  
[References](#Q1-1-43)

## List of Figures

2.1 [Read count density](#x1-13003r1)  
2.2 [Read count sample mean](#x1-13006r2)

## List of Tables

2.1 [Excluded Samples by Batch and Clinical Characteristics.](#x1-13007r1)  
2.2 [Computers used during Thesis](#x1-30001r2)  
2.3 [Linux Utilities and Applications used during Thesis](#x1-30002r3)  
2.4 [R Installations used during Thesis](#x1-31001r4)  
 [2.5 R Packages used during Thesis](#x1-31003)  
2.6 [R Packages Developed during Thesis](#x1-31004r6)

## Glossary

</td></tr><tr

style=“vertical-align:baseline;” id=“TBL-2-2-”>

</td></tr><tr

style=“vertical-align:baseline;” id=“TBL-2-4-”>

synthetic lethal

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

</td></tr></table></div>

## Acronyms

</td></tr><tr

style=“vertical-align:baseline;” id=“TBL-4-2-”>

</td></tr><tr

style=“vertical-align:baseline;” id=“TBL-4-4-”>

ANOVA

Analysis of Variance.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-6-2"

class=“td11”>

BioPAX

Biological Pathway Exchange.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-8-2"

class=“td11”>

CCLE

Cancer Cell Line Encyclopaedia.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-10-2"

class=“td11”>

siRNA

Short interfering ribonucleic acid.

SLIPT

Synthetic lethal interaction prediction tool.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-13-2"

class=“td11”>

TCGA

The Cancer Genome Atlas (genomics project).

</td></tr></table></div>

## Chapter 2 Methods and Resources

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

### 2.1 Bioinformatics Resources for Genomics Research

#### 2.1.1 Public Data and Software Packages

Various bioinformatics resources, such as databases and methods, have become integral parts of genetics and genomics research. Reference genomes, genotyped variants, gene expression, and epigenetics profiles are among the most commonly used resources. Gene expression data in particular is widely available from many microarray and RNA-Seq studies, from repositories such as Gene Expression Omnibus (GEO) ([Clough and Barrett](#XGEO2016), [2016](#XGEO2016)), caArray ([Heiskanen *et al.*](#XcaArray2014), [2014](#XcaArray2014)), and ArrayExpress ([Rustici *et al.*](#XArrayExpress2013), [2013](#XArrayExpress2013)). Such profiles are excellent resources to examine the changes of gene expression occurring in cancers and the variation between samples. These microarray initiatives have set a precedent for data sharing, data mining, and the wider benefits of publicly available data for enabling the scientific community to further utilise the data rather than a single research group or consortium ([Rung and Brazma](#XRung2013), [2013](#XRung2013)). The practice of integrating findings from publicly available genomics data with the research questions and experimental results of individual research groups has carried over into RNA-Seq datasets including the large-scale cancer genomics projects ([Zhang *et al.*](#XICGC2011), [2011](#XICGC2011)). This thesis is one such example of an investigation enabled by this wider movement and tools developed in various disciplines to generate, process, and disseminate genomic-scale data.

Along with databases, it is also becoming common practice for bioinformatics researchers to release their code as open-source or provide a software package to enable replication of the findings or further applications of the methods ([Stajich and Lapp](#XStajich2006), [2006](#XStajich2006)). This is part of a wider movement in software and data analysis with many tools to facilitate such work being released for use in Linux or the R

programming environment ([R Core Team](#XR_core), [2016](#XR_core)). In addition to the R packages hosted on CRAN ([CRAN](#XCRAN), [2017](#XCRAN)), the Bioconductor repositories ([Gentleman *et al.*](#XGentleman2004), [2004](#XGentleman2004)) also contain many packages specifically for applications in bioinformatics, and the GitHub site hosts many packages in various stages of development and early release. Packages from these various sources have been used throughout this project and cited where-ever possible. Several R packages have been developed during this thesis project and either publicly released on GitHub or prepared to accompany a publication.

##### 2.1.1.1 Cancer Genome Atlas Data

Molecular profile data from normal and tumour samples was downloaded from publicly available sources, using the [The Cancer Genome Atlas (TCGA)](#glo:TCGA) ([TCGA](#XTCGA2012), [2012](#XTCGA2012)) and the International Cancer Genome Consortium (ICGC) web portals ([Zhang *et al.*](#XICGC2011), [2011](#XICGC2011)). These include gene expression (RNA-Seq), somatic mutations, and anonymous clinical data. These versions downloaded were on the 6th of August 2015 (Release 19) and the 2nd of May 2016 (Release 20) for breast and stomach cancer respectively via the ICGC data portal (<https://dcc.icgc.org/>).

Performing a genomic alignment in remains a challenge in bioinformatics and methods to do so may yet be improved ([Chen and Tompa](#XChen2010), [2010](#XChen2010)). 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](#glo:TCGA) project ([TCGA](#XTCGA2012), [2012](#XTCGA2012)) used widely adopted tools: “Bowtie” for alignment ([Langmead *et al.*](#Xbowtie), [2009](#Xbowtie)), “mapslice” to detect splice sites ([Wang *et al.*](#Xmapsplice), [2010](#Xmapsplice)), and the Reads Per Kilobase per Million mapped reads (RPKM) approach to qualify reads per transcript as a measure of gene expression ([Mortazavi *et al.*](#XMortazavi2008), [2008](#XMortazavi2008)). These are widely acceptable tools for processing RNA-Seq data which were used to produce the raw counts of mapped reads (tier 1) and normalised expression data (tier 3) publicly available from TCGA.

Raw count and RSEM normalised [TCGA](#glo:TCGA) expression data from Illumina RNA-Seq protocols were available from 1,177 breast samples (113 normal, 1,057 primary tumour, and 7 metastases) for 20,501 genes. [TCGA](#glo:TCGA) breast somatic mutation data for 981 samples (976 primary tumours and 5 metastases) across 25,836 genes were available including 969 samples (964 primary tumours and 5 metastases) with corresponding RNA-Seq expression data and 19,166 genes mapped from Ensembl identifiers to gene symbols, of which 16,156 had

corresponding gene expression information. Unless otherwise stated, the raw counts were used for further processing rather than the RSEM normalised data (provided by [TCGA](#glo:TCGA) tier 3). For the purposes of this analysis somatic mutations were reported if they were detected to non-synonymous substitutions, frameshifts, or truncations (by premature stop codons) which would likely disrupt the wild-type gene function. Normalised protein expression was used (as provided by [TCGA](#glo:TCGA) tier 3), generated from reverse phase protein arrays (RPPA) for 142 antibodies targeting 115 genes for 298 [TCGA](#glo:TCGA) breast samples.

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

Cell line data was downloaded from the [Cancer Cell Line Encyclopaedia (CCLE)](#glo:CCLE) on the 7th of November 2014 ([Barretina *et al.*](#XBarretina2012), [2012](#XBarretina2012); [CCLE](#XCCLE), [2014](#XCCLE)). This includes expression data (gnerated by Affymetrix U133 Plus 2.0 arrays) for 1037 cell lines across 19544 genes (last updated on the 18th of October 2012), DNA copy number, somatic mutation, drug response, and sample information. Samples include 59 breast cell lines and 38 stomach cell lines.

##### 2.1.1.2 Reactome and Annotation Data

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

### 2.2 Data Handling

#### 2.2.1 Normalisation

Apart from the PAM50 subtyping procedure ([Parker *et al.*](#XParker2009), [2009](#XParker2009)), which required RSEM normalised data (J.S. Parker personal communication), the analysis of the RNA-Seq data presented here was based on raw read count data. Raw read counts were log-scaled; samples removed for consistency (based on a Euclidean distance correlation matrix as described in Section [2.2.2](#x1-130002)); and the final dataset was TMM normalised ([Robinson and Oshlack](#XRobinson2010), [2010](#XRobinson2010)) then processed using the voom function ([Law *et al.*](#XLaw2014), [2014](#XLaw2014)) in the limma R package ([Ritchie *et al.*](#Xlimma), [2015](#Xlimma)). Protein expression data generated from RPPA was normalised to dilution curves using the SuperCurve R package ([Ju *et al.*](#XJu2015), [2015](#XJu2015); [Neeley *et al.*](#XNeeley2009), [2009](#XNeeley2009)).

The microarray expression data sourced from the [CCLE](#glo:CCLE) was used as provided, using the Robust Multi-array Average (RMA) and normalized by quantile normalization.

#### 2.2.2 Sample Triage

The [TCGA](#glo: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](#x1-200002r2). 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](#x1-13007r1). These samples showed unusual density plots compared to the rest of the dataset, and exhibited low mean read count in Figures [2.1](#x1-13003r1) and [2.2](#x1-13006r2). 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](#x1-200001r1), and a full correlation heatmap (Figure [A.2](#x1-200002r2)) 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](#x1-2000001)), replicate tumour samples were also examined for sample quality in Appendix [A.2](#x1-2010002). After removal of these samples, the [TCGA](#glo: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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |
| --- |
| (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.

Table 2.1: Excluded Samples by Batch and Clinical Characteristics.

Tissue Source

Type

Batch

Plate

Patient

Samples

p53

</td><td style="white-space:nowrap; text-align:center;" id="TBL-5-1-9"

class=“td11”>

Subtype

<div class="multicolumn" style="white-space:nowrap; text-align:center;"><span

class=“cmbx-12”>Treatment (History)

Clinical Subtypes (Stage)

A7 Christiana

Tumour

47

A227

A0DB

1 of 3

NA

Luminal A

Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-2-10"

class=“td11”> (no)

ER+

Ductal

1. A7 Christiana

* Tumour
* 96
* A220
* A13D
* 1 of 3
* Wildtype
* Luminal A
* Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-3-10"
* class=“td11”> (no)
* ER+
* Ductal
  1. A7 Christiana
  + Tumour
  + 96
  + A227
  + A13E
  + 1 of 3
  + NA
  + Basal
  + Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-4-10"
  + class=“td11”> (no)
  + ER-
  + Ductal
    1. A7 Christiana
    - Tumour
    - 142
    - A277
    - A26E
    - 1 of 3
    - NA
    - Basal
    - Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-5-10"
    - class=“td11”> (no)
    - ER+
    - Ductal
      1. A7 Christiana
      * Tumour
      * 47
      * A277
      * A0DC
      * 1 of 2
      * NA
      * Luminal A
      * Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-6-10"
      * class=“td11”> (yes)
      * ER+
      * Lobular
        1. A7 Christiana
        + Tumour
        + 142
        + A220
        + A26I
        + 1 of 2
        + Mutant
        + Basal
        + Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-7-10"
        + class=“td11”> (yes)
        + ER-
        + Ductal

AC Intl Genomics

Tumour

177

A18M

A2QH

2 of 2

Mutant

Basal

Radical Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-8-10"

class=“td11”> (no)

ER-

Metaplastic

AC Intl Genomics

Tumour

177

A220

A2QH

2 of 2

Mutant

Basal

Radical Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-5-9-10"

class=“td11”> (no)

ER-

Metaplastic

GI ABS IUPUI

Normal

177

A16F

A2C8

1 of 1

NA

Luminal A

Radical Mastectomy and Neoadjuvant

(no)

ER+

Ductal

</td></tr></table> </div>

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

#### 2.2.3 Metagenes and the Singular Value Decomposition

A “metagene” offers a consistent signal of pathway (expression) activation or inactivation by dimension reduction of a matrix, avoiding negatively correlated genes averaging out the signal of a mean-based centroid ([Huang *et al.*](#XHuang2003), [2003](#XHuang2003)). Constructing these pathway metagenes used gene sets for Reactome and Gatza signatures ([Gatza *et al.*](#XGatza2011), [2011](#XGatza2011), [2014](#XGatza2014)) as specified above (see Section [2.1.1.2](#x1-100002)). The singular-value decomposition was performed (X = UT DV where X is the data matrix of the gene set with genes × samples) and the leading eigenvector (first column of V ) corresponding to the largest singular value was used as a metagene for the pathway gene set. To ensure consistent directionality of metagene signals, the median of the gene set in each sample was calculated and correlated against the metagene with the (arbitrary) metagene sign adjusted as needed to conform with the majority of the gene set (i.e., positive correlation between metagene and the median-based centroid). To ensure that genes and pathways were weighted equally, metagenes were derived from a z-transformed dataset of gene expression and samples were scaled (by fractional ranking) for each metagene so that they were comparable on a [0, 1] scale.

##### 2.2.3.1 Candidate Triage and Integration with Screen Data

Candidate triage in combination with the experimental data was intended to integrate findings of the [Synthetic Lethal Interaction Prediction Tool (SLIPT)](#glo:SLIPT) analysis with an ongoing experiment project ([Chen *et al.*](#XChen2014), [2014](#XChen2014); [Telford *et al.*](#XTelford2015), [2015](#XTelford2015)). The first procedure to compare the [SLIPT](#glo:SLIPT) gene candidates for CDH1 with an [short interfering ribonucleic acid (siRNA)](#glo:siRNA) experimental screen ([Telford *et al.*](#XTelford2015), [2015](#XTelford2015)) was a direct comparison of the overlapping candidates, presented in a Venn diagram and tested with the χ2 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.

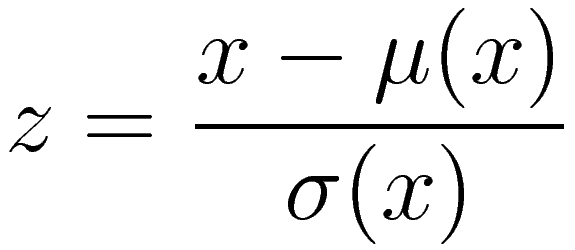
The pathway composition of the intersection was further verified by a permutation resampling analysis (as described in Section [2.3.6](#x1-230006)): the same number of genes detected by [SLIPT](#glo:SLIPT) were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1,652 reactome pathways. These over-representation scores (χ2) were compared the observed over-representation in the intersection of the [SLIPT](#glo:SLIPT) candidates, with the proportion of resamplings with higher χ2 values used for empirical p-values of pathway composition. The χ2 test was used as an appropriation of Fisher’s exact test on a hypergeometric distribution for resampling to computationally scale pathway over-representation tests across iterations. Pathways for which no resamplings were occurred as high as the observed were reported as p < 10-6. These empirical p-values were adjusted for multiple comparisons (FDR). Intersection size was not assumed to be constant across resamplings so similarly with the proportion of resamplings with higher or lower intersection size were used to evaluate significance of enrichment or depletion respectively (of [siRNA](#glo:siRNA) candidate among [SLIPT](#glo: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](#x1-200004) and [2.2.3](#x1-140003), the z-transform has been used to generate z-scores in various analyses in this thesis. Each row of dataset (x) is transformed into a scores (z) using the mean (μ) and standard deviation (σ) of the data such that:



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 3 or -3 respectively.

Empirical Bayes differential expression analysis was performed using the limma R package ([Ritchie *et al.*](#Xlimma), [2015](#Xlimma)). Where specified, the Fisher’s exact test, χ2 test, and correlation were used to measure associations between variables (as implemented in the stats R package ([R Core Team](#XR_core), [2016](#XR_core))). Unless otherwise specified, Pearson’s correlation was used for correlation analyses (r) and coefficient of determination (R2). Where these comparisons are discussed in more detail, Fisher’s exact test and χ2 tests are supported by a table or Venn diagram, rendered with the limma R package ([Ritchie *et al.*](#Xlimma), [2015](#Xlimma)). 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 ([R Core Team](#XR_core), [2016](#XR_core)) has also been used to implement the t-test to compare pairs of data. Where relevant, an [analysis of variance (ANOVA)](#glo:ANOVA) has been performed to report significance of multivariate predictors of outcomes, or least squares linear regression performed for the adjusted coefficient of determination (R2 ) and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violinplot visualisation (rendered in R).

Multiple comparisons are adjusted by the Benjamini-Hochberg procedure to control the false discovery rate (FDR) unless otherwise specified ([Benjamini and Hochberg](#Xfdr1995), [1995](#Xfdr1995)). 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](#XHolm1979), [1979](#XHolm1979)) was also applied in some cases to adjust for multiple comparisons and control the family-wise error rate which adjusts p-values so that the probability that any one of the tests is a false-positive (type-1 error) below a threshold, α.

#### 2.3.2 Gene Set Over-representation Analysis

Gene set enrichment over-representation was performed to test whether there is an enrichment of a gene set (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 [2.3.3](#x1-190003)) or comparison with experimental candidates (see Section [2.2.3.1](#x1-150001)). Initially, these tests were performed using the GeneSetDB web tool ([Araki *et al.*](#Xgenesetdb), [2012](#Xgenesetdb)) hosted by the University of Auckland on the Reactome pathways ([Croft *et al.*](#XReactome), [2014](#XReactome)). Since the GeneSetDB tool used an older version of Reactome (version 40), it was difficult to directly compare with the results of other analysis (see sections [2.2.3.1](#x1-150001) and [2.3.6](#x1-230006)) performed on version 52 (as described in Section [2.1.1.2](#x1-100002)). Thus an implementation of the hypergeometric test in R ([R Core Team](#XR_core), [2016](#XR_core)) 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.*](#Xgenesetdb), [2012](#Xgenesetdb)) were excluded before adjusting for multiple comparisons.

#### 2.3.3 Clustering

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

#### 2.3.4 Heatmap

Standardised z-scores of the data were used to plot heatmaps on an appropriate scale. Raw (log-scale) read counts or voom normalised counts per gene (as specified) were plotted as normalised z-scores on a [-3, +3] blue-red scale. Similarly, correlations were plotted on a [-1, +1] blue-red scale. These heatmaps were performed using the linkage and distance specified for the clustering performed in Section [2.3.3](#x1-190003).

The gplots R package ([Warnes *et al.*](#Xgplots), [2015](#Xgplots)) was used to generate many of the heatmaps throughout this thesis, along with a customised heatmap function (released as heatmap.2x). Where clearly specified, data have been split into subsets with clustering performed separately on each subset with these plotted alongside each other.

#### 2.3.5 Modeling and Simulations

Statistical modeling and simulations have been used to test various synthetic lethal detection procedures on simulated data. This involves constructing a statistical model of how synthetic lethality would appear in (continuous normally distributed) gene expression data. Where presented (in Section 3.2.1), the assumptions of the model are stated clearly. The model allows sampling from a multivariate normal distribution (using the mvtnorm R package ([Genz and Bretz](#XGenz2009), [2009](#XGenz2009); [Genz *et al.*](#Xmvtnorm), [2016](#Xmvtnorm))) to generate simulated data with known underlying synthetic lethal partners (detailed in Section 3.2.2). We can test whether statistical procedures, including those developed in this thesis (presented in Section 3.1), are capable of detecting them upon this simulated data. This multivariate normal simulation procedure also enables the inclusion of correlation structure which is either given as correlated blocks of genes or derived from pathway structures (as detailed in Section 3.4.2).

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

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for χ2, significance and directional criteria met for [SLIPT](#glo: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 discovery

rate” (also used in adjusting for multiple comparisons) was measured here as the proportion of known non-synthetic lethal partners out of all putative partners predicted by the detection procedure. Statistical “accuracy” is the proportion of true predictions for a detection procedure, which is both the correctly predicted known synthetic lethal partners and correctly negative known non-synthetic lethal partners.

##### 2.3.5.1 Receiver Operating Characteristic (Performance)

A more general procedure to measure the statistical performance of a simulation is the Receiver Operating Characteristic (ROC) curve which does not assume a threshold for classification of synthetic lethality but demonstrates the trade-off of sensitivity and specificity ([Akobeng](#XAkobeng2007), [2007](#XAkobeng2007); [Fawcett](#XFawcett2006), [2006](#XFawcett2006); [Zweig and Campbell](#XZweig1993), [1993](#XZweig1993)). These curves (implemented with the ROCR R package ([Sing *et al.*](#XROCR), [2005](#XROCR))) plot the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) as the prediction threshold is varied. An ideal detection method will have a true positive rate of 1 and a false positive rate of 0, hence the Area Under the ROC curve (AUC or AUROC) is a measure of statistical performance for a detection procedure accounting for this trade-off. AUROC values are typically range from 0.5 the value expected by random chance to 1 for an optimal detection method, however it is possible for an AUROC below 0.5 for a poor detection method that performs worse than random chance. An cancer biology, an AUROC of approximately 0.8 is a predictive biomarker suitable for publication ([Hajian-Tilaki](#XHajian-Tilaki2013), [2013](#XHajian-Tilaki2013)) but predictors with lower AUROC values may still be informative depending on the context. In this thesis, the AUROC values varies widely across simulation parameters and a primarily used for comparisons across these parameters, although they can also be used to refine thresholds for optimal classification.

#### 2.3.6 Resampling Analysis

Resampling analyses (e.g., “permutation” analysis) are used to statistically test the significance of an observation without assuming the underlying distribution of expected test statistics [Collingridge](#XCollingridge2013) ([2013](#XCollingridge2013)). Instead these are derived from randomly shuffling test statistics or randomly sampling predicted candidates. For the purposes of this thesis, this involved randomly sampling genes from those tested to be analysed as putative synthetic

lethal candidates. This was performed both for testing the significance of pathway composition in the intersection with experimental gene candidates (Section [2.2.3.1](#x1-150001)) and for assessing the significance of pathway structure among synthetic lethal candidates (Section 3.4.1.1).

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

Resampling was performed for comparison (in Section [2.2.3.1](#x1-150001)) with fixed experimental screen candidates ([Telford *et al.*](#XTelford2015), [2015](#XTelford2015)) both resampling the number of genes overlapping with the screen candidates and test statistics for pathway enrichment. Resampling analysis was also applied to shortest paths and network metrics (in Section 3.4.1.1) to test significance of directional relationships between synthetic lethal candidate genes within pathway structures.

The number of iterations determines the accuracy of these p-values. For pathway composition (in Section [2.2.3.1](#x1-150001)), a million iterations were performed using high performance computing (as detailed in Section [2.5.3](#x1-320003)) 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 signaling, and metabolic pathways ([Barabási and Oltvai](#XBarabasi2004), [2004](#XBarabasi2004)). 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](#Xigraph), [2006](#Xigraph)). Custom R functions to perform more complex analysis and visualisation of iGraph data objects will be described later.

Graph theory is a branch of pure mathematics which deals with the properties of sets of discrete objects (referred to as a ‘node’ or ‘vertex‘) with some pairs are joined (by a ‘link’ or an ‘edge‘). While a seemingly reductionist abstraction to mathematically study relationships, graph theory serves has applications in a wide range of studies including life sciences. Network theory is the sub-discipline of graph theory which deals with networks which has become popular due to the vast potential for applications of networks ([van Steen](#XvanSteen2010), [2010](#XvanSteen2010)).

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](#XvanSteen2010), [2010](#XvanSteen2010)). Unless otherwise stated, graph structures and networks in thesis will be unweighted and have no parallel edges or loops. Where a directional relationship is known or modelled, it will be represented with a directed edge in a directed graph.

#### 2.4.2 Sourcing Graph Structure Data

Pathway Commons interaction data was sourced using [Biological pathway exchange (BioPAX)](#glo:BioPAX) with the paxtools-4.3.0 Java application on October 6th 2015 ([Cerami *et al.*](#XPathwayCommons), [2011](#XPathwayCommons); [Demir *et al.*](#Xpaxtools), [2013](#Xpaxtools)). This utility was used to source ‘sif’ format interaction data into R ([R Core Team](#XR_core), [2016](#XR_core)), from which the human Reactome (version 52) dataset of interactions was imported ([Croft *et al.*](#XReactome), [2014](#XReactome)), 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](#x1-100002)). A subgraph with adjacent nodes was constructed by adding nodes which have an edge with a gene in the pathway gene set. The pathways these adjacent nodes belong to were added to form a “meta-pathway” to account for the possibility for nodes within the pathway being linked by the surrounding graph structure.

#### 2.4.4 Network Analysis Metrics

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

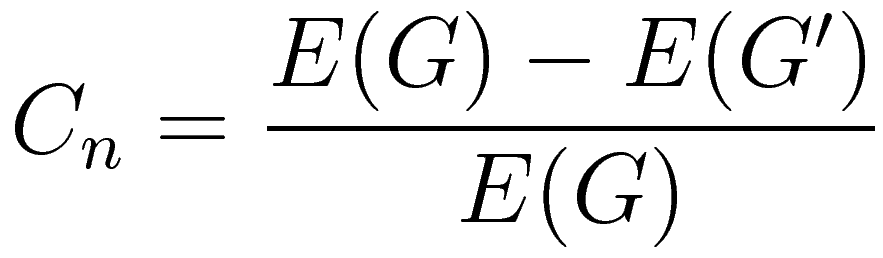
Vertex degree is the number of edges a node has and is a fundamental measure of the importance and connectivity of a network ([van Steen](#XvanSteen2010), [2010](#XvanSteen2010)). More connected nodes, such as network hubs, will have a higher vertex degree relative to other nodes. For the purposes of 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](#XDijkstra1959), [1959](#XDijkstra1959)) in the igraph R package ([Csardi and Nepusz](#Xigraph), [2006](#Xigraph)). 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).

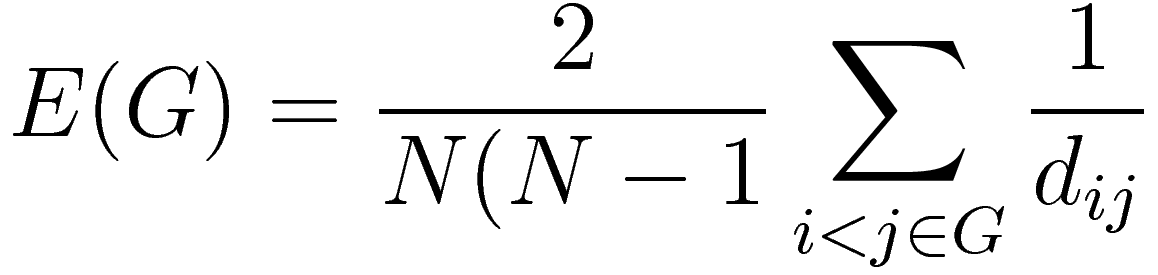
Network centrality is an alternative measure of the importance or influence of a node to the graph structure ([Borgatti](#XBorgatti2005), [2005](#XBorgatti2005)). Various strategies are used to derive centrality, typically based on how connected the node is or the impact of node removal on the connectivity 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](#XBrin1998), [1998](#XBrin1998)). This is implemented in the igraph R package ([Csardi and Nepusz](#Xigraph), [2006](#Xigraph)).

Another network centrality measure that has been previously applied to biological protein interaction networks ([Kranthi *et al.*](#XKranthi2013), [2013](#XKranthi2013)) is the “information centrality”. The information centrality of a node is the relative impact on efficiency (transmission of information via shortest paths) of the network when the node is removed. That is the centrality (C) ([Kranthi *et al.*](#XKranthi2013), [2013](#XKranthi2013)) for node n in graph G is defined as:



where G′ is the subgraph with the node removed and E is the efficiency ([Latora and Marchiori](#XLatora2001), [2001](#XLatora2001)) derived from shortest paths (dij between nodes i and j) 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.

### 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](#x1-30001r2)), 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](#x1-30001r2)) and interacting with these via ZSH shell. However, Ubuntu OS and the Bourne Again SHell (bash) were used at the inception of this project and bash is continues to be used for running scripts. Various Linux applications and command-line utilities were used on these machines (as summarised in Table [2.3](#x1-30002r3)). 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

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-6-1-2"

class=“td11”>

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-6-1-3"

class=“td11”>

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-6-1-4"

class=“td11”>

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-6-1-5"

class=“td11”>

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 |

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Upstream OS</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-6-4-2"

class=“td11”>

|  |
| --- |
| Ubuntu LTS |
| Trusty 14.04 |

|  |
| --- |
| Ubuntu LTS |
| Xenial 16.04 |

</td><td style="white-space:nowrap; text-align:left;" id="TBL-6-4-5"

class=“td11”>

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Linux Kernel</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-6-5-2"

class=“td11”> 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

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Shell: bash</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-6-6-2"

class=“td11”> 4.3.11(1)

4.3.46(1)

4.2.46(1)

4.2.1(1)

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Shell: zsh</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-6-7-2"

class=“td11”> 5.0.2

5.1.1

5.0.2

5.2

</td></tr></table> </div>

Table 2.3: Linux Utilities and Applications used during Thesis

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-1-2"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-1-3"

class=“td11”>

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-1-4"

class=“td11”>

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-1-5"

class=“td11”>

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-1-6"

class=“td11”>

NeSI Pan Cluster

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-3-2"

class=“td11”>

OS

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-3-3"

class=“td11”>

|  |
| --- |
| Elementary OS |
| Freya 0.3.2 |

|  |
| --- |
| Elementary OS |
| Loki 0.4 |

|  |
| --- |
| Red Hat Enterprise |
| Maipo 7.2 |

|  |
| --- |
| Cent OS |
| Final 6.4 |

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-4-2"

class=“td11”>

Linux Kernel

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-4-3"

class=“td11”> 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

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-5-3"

class=“td11”> 4.3.11(1)

4.3.46(1)

4.2.46(1)

4.2.1(1)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-6-2"

class=“td11”>

Shell zsh

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-6-3"

class=“td11”> 5.0.2

5.1.1

5.0.2

5.2

Programming

Python

2.7.6

2.7.12

2.7.5

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-8-2"

class=“td11”>

Java

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-8-3"

class=“td11”> 1.8.0\_101

9-ea

1.8.0\_101

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-9-2"

class=“td11”>

C++

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-9-3"

class=“td11”> 4.8.4

5.4.0

4.8.5

4.4.7

Text Editor

nano

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-10-3"

class=“td11”> 2.2.6

2.5.3

2.3.1

2.0.9

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-11-2"

class=“td11”>

kile (LATEX)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-11-3"

class=“td11”> 2.1.3

2.1.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-11-6"

class=“td11”>

Version Control

git

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-12-3"

class=“td11”> 1.9.1

2.11.0

1.7.1

1.8.3.1

Shell Utilities

sed

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-13-3"

class=“td11”> 4.4.2

4.4.2

4.4.2

4.4.1

grep

2.16-1

2.25-1

2.20

2.6.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-15-2"

class=“td11”>

nohup

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-15-3"

class=“td11”> 8.21

8.25

8.22

8.4

Typesetting

TEX

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-16-3"

class=“td11”> 3.1415926

3.14159265

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-16-6"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-17-2"

class=“td11”>

TexLive (LATEX)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-17-3"

class=“td11”> 2013

2015

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-17-6"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-18-2"

class=“td11”>

PDFTEX

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-18-3"

class=“td11”> 2.5-1

2.6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-18-6"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-19-2"

class=“td11”>

pandoc

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-19-3"

class=“td11”> 1.12.2.1

1.16.0.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-19-6"

class=“td11”>

Remote Computing

slurm scheduler

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-20-3"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-20-5"

class=“td11”>

16.05.6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-21-2"

class=“td11”>

OpenSSH

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-21-3"

class=“td11”> 7.2p2

7.2p2

6.6.1

5.3p1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-22-2"

class=“td11”>

OpenSSL

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-22-3"

class=“td11”> 1.0.2g

1.0.2g

1.0.01e-fips

1.0.01e-fips

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-23-2"

class=“td11”>

rsync

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-23-3"

class=“td11”> 3.1.0p31

3.1.1p31

3.0.9p30

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-24-2"

class=“td11”>

Globus Online Transfer

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-24-4"

class=“td11”>

3.1

3.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-25-2"

class=“td11”>

Cisco AnyConnect VPN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-25-4"

class=“td11”> 3.1.05170

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-25-6"

class=“td11”>

Image Processing

Inkscape

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-26-3"

class=“td11”> 0.48.4

0.91

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-26-6"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-27-2"

class=“td11”>

GIMP

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-27-3"

class=“td11”> 2.8.10

2.8.16

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-27-6"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-28-2"

class=“td11”>

ImageMagick

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-28-3"

class=“td11”> 6.7.7.10-6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-13-28-5"

class=“td11”>

</td>

</td></tr></table> </div>

#### 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](#x1-31001r4). 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](#x1-31002r5) with versions specified) were installed from the Comprehensive R Archive Network ([CRAN](#XCRAN), [2017](#XCRAN)), Bioconductor ([Gentleman *et al.*](#XGentleman2004), [2004](#XGentleman2004), version 3.4; BiocInstaller 1.24.0), or GitHub. These packages were not updated when they would change the functionality of scripts or functions in packages, in particular imported data from annotation packages (used to define gene sets) have been saved as local files to continue using stable versions of these pathway data (across machines).

This is a summary of the key packages which (in addition to their dependencies) have been used throughout this project. Where a package implementation has been central to the methods applied, they are described in more detail in the relevant section. A full table of packages used in this thesis can be found in Appendix [B](#x1-202000B) (Table [B.1](#x1-202001r1)). The R packages developed during this thesis are given in Table [2.6](#x1-31004r6) 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

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-1-2"

class=“td11”>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-1-3"

class=“td11”>

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-1-4"

class=“td11”>

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-1-5"

class=“td11”>

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-1-6"

class=“td11”>

NeSI Pan Cluster

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-3-2"

class=“td11”>

OS

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-3-3"

class=“td11”>

|  |
| --- |
| Elementary OS |
| Freya 0.3.2 |

|  |
| --- |
| Elementary OS |
| Loki 0.4 |

|  |
| --- |
| Red Hat Enterprise |
| Maipo 7.2 |

|  |
| --- |
| Cent OS |
| Final 6.4 |

Programming

R

</td><td style="white-space:nowrap; text-align:left;" id="TBL-18-4-3"

class=“td11”> 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)

</td>

</td></tr></table> </div>

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

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-1-2"

class=“td11”>

Package Name

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-1-3"

class=“td11”>

Description and GitHub Repository

</td><td style="white-space:nowrap; text-align:center;" id="TBL-24-1-4"

class=“td11”>

Section

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-3-2"

class=“td11”>

slipt

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-3-3"

class=“td11”>

|  |
| --- |
| Synthetic lethal detection by [SLIPT](#glo:SLIPT) (to accompany publication) |
| <https://github.com/TomKellyGenetics/slipt> |

3.1

visualisation

vioplotx

|  |
| --- |
| Customised violin plots (based on vioplot) |
| <https://github.com/TomKellyGenetics/vioplotx> |

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-5-2"

class=“td11”> heatmap.2x

|  |
| --- |
| Customised heatmaps (based on gplots) |
| <https://github.com/TomKellyGenetics/heatmap.2x> |

[2.3.4](#x1-200004)

igraph.extensions

igraph.extensions

|  |
| --- |
| Meta-package to install the follow iGraph functions |
| <https://github.com/TomKellyGenetics/igraph.extensions> |

3.5.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-7-2"

class=“td11”> plot.igraph

|  |
| --- |
| Custom plotting of directed graphs |
| <https://github.com/TomKellyGenetics/plot.igraph> |

[2.4.4](#x1-280004)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-8-2"

class=“td11”> info.centrality

|  |
| --- |
| Computing information centrality from network efficiency |
| <https://github.com/TomKellyGenetics/info.centrality> |

3.4.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-9-2"

class=“td11”> pathway.structure.permutation

|  |
| --- |
| Testing pathway structure with resampling analysis |
| <https://github.com/TomKellyGenetics/pathway.structure.permutation> |

3.4.1.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-24-10-2"

class=“td11”> graphsim

|  |
| --- |
| Generating simulated expression from graph structures |
| <https://github.com/TomKellyGenetics/graphsim> |

3.4.2

</td></tr></table> </div>

#### 2.5.3 High Performance and Parallel Computing

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

The New Zealand eScience Infrastructure (NeSI) is a computating resource providing the Intel Pan cluster hosted by the University of Auckland ([NeSI](#XNeSI), [2017](#XNeSI)). 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](#x1-230006) and 3.4.1.1), calculating information centrality (in Section [2.4.4](#x1-280004)), and in simulations (in sections [2.3.5](#x1-210005), 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](#XGlobus), [2017](#XGlobus)). R scripts ([R Core Team](#XR_core), [2016](#XR_core)) were run in parallel with the “simple network of workstations” snow R package [Tierney *et al.*](#Xsnow) ([2015](#Xsnow)). This utilised the “message passing interface” ([Yu](#XRmpi), [2002](#XRmpi)) 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 available resources and run on the Pan cluster via the Slurm (Simple Linux Utility for Resource Management) workload manager ([Slurm](#Xslurm), [2017](#Xslurm)). 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.

## References

   Aarts, M., Bajrami, I., Herrera-Abreu, M.T., Elliott, R., Brough, R., Ashworth, A., Lord, C.J., and Turner, N.C. (2015) Functional genetic screen identifies increased sensitivity to wee1 inhibition in cells with defects in fanconi anemia and hr pathways. *Mol Cancer Ther*, **14**(4): 865–76.

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

   Adamski, M.G., Gumann, P., and Baird, A.E. (2014) A method for quantitative analysis of standard and high-throughput qPCR expression data based on input sample quantity. *PLoS ONE*, **9**(8): e103917.

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

   Agarwal, S., Deane, C.M., Porter, M.A., and Jones, N.S. (2010) Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol*, **6**(6): e1000817.

   Agrawal, N., Akbani, R., Aksoy, B.A., Ally, A., Arachchi, H., Asa, S.L., Auman, J.T., Balasundaram, M., Balu, S., Baylin, S.B., *et al.* (2014) Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**(3): 676–690.

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

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

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

   American Society for Clinical Oncology (ASCO) (2017) The genetics of cancer. [http://www.cancer.net/navigating-\_cancer-\_care/cancer-\_basics/genetics/genetics-\_cancer](http://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer). Accessed: 22/03/2017.

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

   Araki, H., Knapp, C., Tsai, P., and Print, C. (2012) GeneSetDB: A comprehensive

meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**: 76–82.

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

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

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

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

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

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

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

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

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

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

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

   Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.Y., Ou, J., San Luis, B.J., Bandyopadhyay, S., *et al.* (2010b) Quantitative analysis

of fitness and genetic interactions in yeast on a genome scale. *Nat Meth*, **7**(12): 1017–1024.

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

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

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

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

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

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

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

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

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

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

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

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

   Boettcher, M., Lawson, A., Ladenburger, V., Fredebohm, J., Wolf, J., Hoheisel, J.D., Frezza, C., and Shlomi, T. (2014) High throughput synthetic lethality screen reveals a tumorigenic role of adenylate cyclase in fumarate hydratase-deficient cancer cells. *BMC Genomics*, **15**: 158.

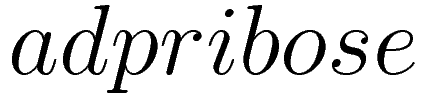
   Boone, C., Bussey, H., and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, **8**(6): 437–49.

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

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

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

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

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

   Burk, R.D., Chen, Z., Saller, C., Tarvin, K., Carvalho, A.L., Scapulatempo-Neto, C., Silveira, H.C., Fregnani, J.H., Creighton, C.J., Anderson, M.L., *et al.* (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**(7645): 378–384.

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

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

   Cancer Research UK (2017) Family history and cancer genes. [http://www.cancerresearchuk.org/about-\_cancer/causes-\_of-\_cancer/inherited-\_cancer-\_genes-\_and-\_increased-\_cancer-\_risk/family-\_history-\_and-\_inherited-\_cancer-\_genes](http://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/family-history-and-inherited-cancer-genes). Accessed: 22/03/2017.

   Cancer Cell Line Encyclopedia (CCLE) (2014) Broad-Novartis Cancer Cell Line Encyclopedia. <http://www.broadinstitute.org/ccle>. Accessed: 07/11/2014.

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

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

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

   Chen, K., Yang, D., Li, X., Sun, B., Song, F., Cao, W., Brat, D.J., Gao, Z., Li, H., Liang, H., *et al.* (2015) Mutational landscape of gastric adenocarcinoma in Chinese: implications for prognosis and therapy. *Proc Natl Acad Sci USA*, **112**(4): 1107–1112.

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

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

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

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

   Christofori, G. and Semb, H. (1999) The role of the cell-adhesion molecule

E-cadherin as a tumour-suppressor gene. *Trends in Biochemical Sciences*, **24**(2): 73 – 76.

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

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

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

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

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

   Collins, F.S., Morgan, M., and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, **300**(5617): 286–290.

   Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer,

D., Cope, L., Creighton, C., Danilova, L., *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**(7511): 543–550.

   Corcoran, R.B., Ebi, H., Turke, A.B., Coffee, E.M., Nishino, M., Cogdill, A.P., Brown, R.D., Della Pelle, P., Dias-Santagata, D., Hung, K.E., *et al.* (2012) Egfr-mediated reactivation of mapk signaling contributes to insensitivity of BRAF -mutant colorectal cancers to raf inhibition with vemurafenib. *Cancer Discovery*, **2**(3): 227–235.

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

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

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

   Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res*, **42**(database issue): D472D477.

   Crunkhorn, S. (2014) Cancer: Predicting synthetic lethal interactions. *Nat Rev Drug Discov*, **13**(11): 812.

   Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal*, **Complex Systems**: 1695.

   Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403): 346–352.

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

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

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

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

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

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

   Dienstmann, R. and Tabernero, J. (2011) BRAF as a target for cancer therapy. *Anticancer Agents Med Chem*, **11**(3): 285–95.

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

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

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

   Dorogovtsev, S.N. and Mendes, J.F. (2003) *Evolution of networks: From biological nets to the Internet and WWW* . Oxford University Press, USA.

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

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

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

   Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M.L. (2014) Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, **23**(22): 5866.

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

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

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

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

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

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

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

   Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., *et al.* (2015) COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**(Database issue): D805–811.

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

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

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

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

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

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

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

   Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S.,

Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10): R80.

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

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

   Gilbert, W. and Maxam, A. (1973) The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences*, **70**(12): 3581–3584.

   Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P., and Caldas, C. (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, **16**(5): 991–1006.

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

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

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

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

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

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

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

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

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

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

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

   Han, J.D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**(6995): 88–93.

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

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

   Hanna, S. (2003) Cancer incidence in new zealand (2003-2007). In D. Forman, D. Bray F Brewster, C. Gombe Mbalawa, B. Kohler, M. Piñeros,

E. Steliarova-Foucher, R. Swaminathan, and J. Ferlay (editors), *Cancer Incidence in Five Continents*, volume X, 902–907. International Agency for Research on Cancer, Lyon, France. Electronic version <http://ci5.iarc.fr> Accessed 22/03/2017.

   Heiskanen, M., Bian, X., Swan, D., and Basu, A. (2014) caArray microarray database in the cancer biomedical informatics gridTM (caBIGTM). *Cancer Research*, **67**(9 Supplement): 3712–3712.

   Heiskanen, M.A. and Aittokallio, T. (2012) Mining high-throughput screens for cancer drug targets-lessons from yeast chemical-genomic profiling and synthetic lethality. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**(3): 263–272.

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

   Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., Chandrasekharan, S., *et al.* (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol*, **8**(5): R76.

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

   Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng,

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

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

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

   Holme, P. and Kim, B.J. (2002) Growing scale-free networks with tunable clustering. *Physical Review E*, **65**(2): 026107.

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

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

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

   Illumina, Inc (Illumina) (2017) Sequencing and array-based solutions for genetic research. <https://www.illumina.com/>. Accessed: 26/03/2017.

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

   Internationl Human Genome Sequencing Consortium (IHGSC) (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011): 931–945.

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

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

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

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

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

   Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., Yamamoto, S., Tatsuno, K., Katoh, H., Watanabe, Y., *et al.* (2014) Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. *Nat Genet*, **46**(6): 583–587.

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

   Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., Benz, C.C., *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**(7447): 67–73.

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

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

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

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

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

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

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

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

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

   Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and

memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**(3): R25.

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

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

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

   Lawrence, M.S., Sougnez, C., Lichtenstein, L., Cibulskis, K., Lander, E., Gabriel, S.B., Getz, G., Ally, A., Balasundaram, M., Birol, I., *et al.* (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**(7536): 576–582.

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

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

   Lee, A.Y., Perreault, R., Harel, S., Boulier, E.L., Suderman, M., Hallett, M., and Jenna, S. (2010a) Searching for signaling balance through the identification of genetic interactors of the rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS ONE*, **5**(5): e10624.

   Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G., and Marcotte, E.M. (2010b) Predicting genetic modifier loci using functional gene networks. *Genome Research*, **20**(8): 1143–1153.

   Lee, I. and Marcotte, E.M. (2009) Effects of functional bias on supervised learning of a gene network model. *Methods Mol Biol*, **541**: 463–75.

   Lee, M.J., Ye, A.S., Gardino, A.K., Heijink, A.M., Sorger, P.K., MacBeath, G., and Yaffe, M.B. (2012) Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell*, **149**(4): 780–94.

   Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A.G. (2006) Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, **38**(8): 896–903.

   Li, X.J., Mishra, S.K., Wu, M., Zhang, F., and Zheng, J. (2014) Syn-lethality: An integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. *Biomed Res Int*, **2014**: 196034.

   Linehan, W.M., Spellman, P.T., Ricketts, C.J., Creighton, C.J., Fei, S.S., Davis, C.,

Wheeler, D.A., Murray, B.A., Schmidt, L., Vocke, C.D., *et al.* (2016) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med*, **374**(2): 135–145.

   Lokody, I. (2014) Computational modelling: A computational crystal ball. *Nature Reviews Cancer*, **14**(10): 649–649.

   Lord, C.J., Tutt, A.N., and Ashworth, A. (2015) Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annu Rev Med*, **66**: 455–470.

   Lu, X., Kensche, P.R., Huynen, M.A., and Notebaart, R.A. (2013) Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nat Commun*, **4**: 2124.

   Lu, X., Megchelenbrink, W., Notebaart, R.A., and Huynen, M.A. (2015) Predicting human genetic interactions from cancer genome evolution. *PLoS One*, **10**(5): e0125795.

   Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., Wolf, M.K., Butler, J.S., Hinshaw, J.C., Garnier, P., Prestwich, G.D., Leonardson, A., *et al.* (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**(1): 121–137.

   Luo, J., Solimini, N.L., and Elledge, S.J. (2009) Principles of Cancer Therapy: Oncogene and Non-oncogene Addiction. *Cell*, **136**(5): 823–837.

   Machado, J., Olivera, C., Carvalh, R., Soares, P., Berx, G., Caldas, C., Sercuca, R., Carneiro, F., and Sorbrinho-Simoes, M. (2001) E-cadherin gene (CDH1) promoter methylation as the second hit in sporadic diffuse gastric carcinoma. *Oncogene*, **20**: 1525–1528.

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

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

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

   McCourt, C.M., McArt, D.G., Mills, K., Catherwood, M.A., Maxwell, P., Waugh, D.J., Hamilton, P., O’Sullivan, J.M., and Salto-Tellez, M. (2013) Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. *PLoS ONE*, **8**(7): e69604.

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

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

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

   Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7): 621–628.

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

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

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

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

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

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

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

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

   Peltonen, L. and McKusick, V.A. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, **291**(5507): 1224–1229.

   Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., *et al.* (2016) Erratum: The somatic

mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*, **7**: 11908.

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

   Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R., *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**(7278): 191–196.

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

   Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R.L., Bardelli, A., and Bernards, R. (2012) Unresponsiveness of colon cancer to BRAF(v600e) inhibition through feedback activation of egfr. *Nature*, **483**(7387): 100–3.

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

   Ravnan, M.C. and Matalka, M.S. (2012) Vemurafenib in patients with BRAF v600e mutation-positive advanced melanoma. *Clin Ther*, **34**(7): 1474–86.

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

   Robin, J.D., Ludlow, A.T., LaRanger, R., Wright, W.E., and Shay, J.W. (2016) Comparison of DNA Quantification Methods for Next Generation Sequencing. *Sci Rep*, **6**: 24067.

   Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**(3): R25.

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

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

   Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., *et al.* (2013) ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res*, **41**(Database issue): D987–990.

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

   Sander, J.D. and Joung, J.K. (2014) Crispr-cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, **32**(4): 347–55.

   Sanger, F. and Coulson, A. (1975) A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, **94**(3): 441 – 448.

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

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

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

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

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

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

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

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

   Sun, C., Wang, L., Huang, S., Heynen, G.J.J.E., Prahallad, A., Robert, C., Haanen, J., Blank, C., Wesseling, J., Willems, S.M., *et al.* (2014) Reversible and adaptive resistance to BRAF(v600e) inhibition in melanoma. *Nature*, **508**(7494): 118–122.

   Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*, **27**(2): 199–204.

   Telford, B.J., Chen, A., Beetham, H., Frick, J., Brew, T.P., Gould, C.M., Single, A.,

Godwin, T., Simpson, K.J., and Guilford, P. (2015) Synthetic lethal screens identify vulnerabilities in gpcr signalling and cytoskeletal organization in E-cadherin-deficient cells. *Mol Cancer Ther*, **14**(5): 1213–1223.

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

   The Cancer Genome Atlas Research Network (TCGA) (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418): 61–70.

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

   The Cancer Genome Atlas Research Network (TCGA) (2017b) The Cancer Genome Atlas Project Data Portal. [https://tcga-\_data.nci.nih.gov/](https://tcga-data.nci.nih.gov/). Accessed: 06/02/2017 (via cBioPortal.

   The Cancer Society of New Zealand (Cancer Society of NZ) (2017) What is cancer? [https://otago-\_southland.cancernz.org.nz/en/cancer-\_information/other-\_links/what-\_is-\_cancer-\_3/](https://otago-southland.cancernz.org.nz/en/cancer-information/other-links/what-is-cancer-3/). Accessed: 22/03/2017.

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

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

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

   The Internation Cancer Genome Consortium (ICGC) (2017) ICGC Data Portal. <https://dcc.icgc.org/>. Accessed: 06/02/2017.

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

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

   The Pharmaceutical Management Agency (PHARMAC) (2016) Approval of multi-product funding proposal with roche.

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

   Tiong, K.L., Chang, K.C., Yeh, K.T., Liu, T.Y., Wu, J.H., Hsieh, P.H., Lin, S.H.,

Lai, W.Y., Hsu, Y.C., Chen, J.Y., *et al.* (2014) Csnk1e/ctnnb1 are synthetic lethal to tp53 in colorectal cancer and are markers for prognosis. *Neoplasia*, **16**(5): 441–50.

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

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

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

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

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

   Tsai, H.C., Li, H., Van Neste, L., Cai, Y., Robert, C., Rassool, F.V., Shin, J.J., Harbom, K.M., Beaty, R., Pappou, E., *et al.* (2012) Transient low doses of dna-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell*, **21**(3): 430–46.

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

   van der Meer, R., Song, H.Y., Park, S.H., Abdulkadir, S.A., and Roh, M. (2014) RNAi screen identifies a synthetic lethal interaction between PIM1 overexpression and PLK1 inhibition. *Clinical Cancer Research*, **20**(12): 3211–3221.

   van Steen, K. (2012) Travelling the world of genegene interactions. *Briefings in Bioinformatics*, **13**(1): 1–19.

   van Steen, M. (2010) *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen, VU Amsterdam.

   Vapnik, V.N. (1995) *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

   Vargas, J.J., Gusella, G., Najfeld, V., Klotman, M., and Cara, A. (2004) Novel integrase-defective lentiviral episomal vectors for gene transfer. *Hum Gene Ther*, **15**: 361–372.

   Vizeacoumar, F.J., Arnold, R., Vizeacoumar, F.S., Chandrashekhar, M., Buzina,

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

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

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

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

   Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., *et al.* (2014) Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet*, **46**(6): 573–582.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

   Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., *et al.* (2011) International cancer genome

consortium data 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.