Library Declaration Form

University of Otago Library

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

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

Degree: Doctor of Philosophy

Department: Department of Biochemistry

Permanent Address: 710 Cumberland Street, Dunedin, NZ

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

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

i) a library

ii) an individual

at the discretion of the University of Otago.

Signature:

Date:

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

S. Thomas Kelly

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand.

31 March 2017

Abstract

Background

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

Approach

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

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

Findings

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

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

of directionality between the candidates was detected.

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

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

Acknowledgements

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

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

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

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

I must also thank various organisations supported this research project:

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

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

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

Thanks most of all to my fianceé, Dr Yui Kawagishi, you've been an inspiration. Thank you for your support, help, and encouragement, even from afar times, it has always made a difference. It's been incredible to see you flourish in your career and I look forward to joining you again soon. May the next chapter of our adventures involve a bit less Skype across timezones.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The thesis presents research into genetic interactions based on genomics data and bioinformatics approaches. This chapter introduces the recent developments in genomics and bioinformatics, particularly in their application to cancer research. Synthetic lethal interactions are a long standing area of research in genetics in both model organisms and cancer biology. Various reasons why these interactions are of interest in fundamental and translational biology will be outlined but first these and similar interactions will be defined. A bioinformatics approach to synthetic lethal interactions enables much wider exploration of the inter-connected nature of genes and proteins within a cancer cell than previous candidate-based approaches. An alternative approach is experimental screening which will be presented and contrasted with bioinformatics approaches in more detail. An emerging application of synthetic lethality is the design of treatments with specificity against loss of function mutations in tumour suppressor genes. E-cadherin (encoded by *CDH1*) is a prime example of this which will be the focus of the analysis in this thesis and as such the role of this gene in cellular and cancer biology will be briefly reviewed.

1.1 Cancer Research in the Post-Genomic Era

Genomics technologies have the potential to vastly impact upon various areas including health and cancer medicine. Considering the progress in recent genomics research, it could soon impact greatly upon clinical and wider applications of genetics either directly or by enabling more focused genetics research from candidates selected from genomics or bioinformatics analysis. The completion of the draft Human Genome (Lander *et al.*, 2001) marks a major accomplishment in genetics research and raises

new challenges to utilise this genomic scale information effectively. Technologies in this area have rapidly developed since completion of the human genome project and many global large-scale projects have to expanded upon the human genome, to populations (1000 Genomes, 2010), to cancers (Dickson, 1999; Zhang *et al.*, 2011), and to deeper functional understanding (Kawai *et al.*, 2001; ENCODE, 2004). However, impact on the clinic has been slower than initially anticipated following the completion of the "draft" genome with genomics technologies yet to become widely adopted in healthcare and oncology. Here we outline the genomics technologies and bioinformatics approaches which have led to availability of genomics data and techiques used in this thesis and potential for applications in cancer research or the clinic in the future.

1.1.1 Cancer as a Global Health Concern

Cancer is a class of diseases involving malignant cellular growth, invasion of tissues, and spread to other organs. While there are also environmental factors, most cancers occur more frequently with age and family history so genetics is widely acknowledged to have an important role in cancer risk. Cancers arise from dyregulated cellular growth or differentiation from stem cells, these can occur through genetic mutations or alterations in gene regulation or expression.

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

While the genetic contribution to cancer risk and many of the molecular changes occuring cancers are widely acknowledged (ASCO, 2017; Cancer Research UK, 2017; Cancer Society of NZ, 2017), much of these findings have yet to impact on clinical practice. Diagnostics are traditionally based on pathological examination of cancer cell and tissue samples, including histological staining for biomolecules and biomarkers, and continue to be widely used. The current standard of care is surgery, radiation, and

cytotoxic chemotherapy, depending on whether the cancer is localised or has become systemic (via metastasis) and spread to other organ systems. These approaches are effective against cancers, particularly in patients particular subtypes (such as acute myeloid leukaemia) or early stage cancers. Thus early intervention is important to patient survival and quality of life with national screening programs aiming to diagnose cancers early and subtypes more accurately, including identification of patients with genetic variants or family histories for high risk of particular cancers.

Chemotherapy is a treatment for advanced stage (systemic) cancers which is designed to inhibit the growth and spread of cancer throughout the body by targeting rapidly growing cells. However, this approach is notorious for adverse effects and a narrow therapeutic window and is not suitable for chemopreventative application in many cases (Kaelin, Jr, 2009). Thus high risk individuals are regularly monitored for cancers and offered preventative surgery (Guilford *et al.*, 2010; Scheuer *et al.*, 2002), although this is not completely effective at preventing cancers and may impact on quality of life, depending on the cancer tissue types they are at risk of. Alternative treatment strategies based on molecular biology and other fields are being investigated, including immunological, endocrine, and targeted therapeutics, with a particular interest in treatments with specificity against cancer cells and wider applications (i.e., tolerable effective doses in applciations as a chemopreventative or against advanced stage cancers).

1.1.1.1 Genetics and Molecular Biology in Cancers

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

Hanahan and Weinberg (2000) identified several key molecular and cellular traits shared across most cancers as a rational approach to the complex change that occur cancer initiation and progresion due to common molecular machinery underlying all cells. A cancer cell must posess limitless replication potential, modulate growth signals to grow indefinitely, and gain invasive or metastatic capabilities. In addition, cancers must evade apoptosis, the immune system, and sustain angiogenesis and enery metabolism in order to survive (Hanahan and Weinberg, 2000, 2011). In order to achieve this, cancer cells undergo changes to their genomes and the surrounding cells to create a tumour microenvironment. Thus genomic instability has a key role in the survival and proliferation of cancer cells and the progression of further disease, as these malignant characteristics are acquired. Identifying the mechanisms of these acquired traits and the underlying genetic mutation or dysregulation behind them, such as E-cadherin mutation in metastasis or p53 mutation in genomic instability (Hanahan and Weinberg, 2000), will be an important step in understanding and inhibiting cancer with the next generation of genomically-informed treatments.

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

prognostic importance for patients Dai *et al.* (2015). High-throughput technologies have the potential to enable such subtyping on a vast scale in discovery of further subtypes in breast cancer or other diseases and in identification of these subtypes along with mutations in routine clinical diagnostic and prognostic testing. The "Pan cancer" approaches by the cancer genome atlas project (as discussed in more detail in section 1.1.5.2.1) expand on the importance of molecular differences between cancers by examining molecular profiles across cancer tissue types (Weinstein *et al.*, 2013).

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

1.1.2 The Human Genome Revolution

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

1.1.2.1 The First Human Genome Sequence

The first human genome is a good example of a large-scale genomics project for it's success as an international collaboration and releasing their data as a resource for the wider scientific community (Collins *et al.*, 2003; Lander *et al.*, 2001). This particular project generated significant public interest due to it being a landmark achievement,

the first of it's scale, and some controversial findings. Namely, the number of genes discovered (particularly those specific to vertbrates) was much lower than most estimates of a genome of it's size and the number of repetious transposon elements was very high. Even the figure of 30–40,000 genes given by the original publication is now regarded to be an overestimate (Ezkurdia *et al.*, 2014; IHGSC, 2004).

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

The "whole genome shotgun" approach (now widely used in genomics sequencing) was pioneered by a competing private genome project completed shortly afterwards by Celera Genomics, demonstrating the power and speed of this approach by sequencing 27 million reads of the entire 2.91Gbp human genome (5.11× coverage) in only 9-months (Venter *et al.*, 2001). Assembly was assisted with the 2.9× coverage public genome data, reduced to raw shotgun reads to reomve cloning bias. While, repetious sequences remained an issue for this project, more than 90% of the genome was able to be assembled into 100kbp scaffolds and 26,588 protein coding genes were identified, closer to the current consensus for the number of genes in the human genome. This project in particular emphasised the value of computational assembly methods in handling a large number of reads, reducing the time and cost of sequencing, and established the shotgun approach for wider adoption with more recent sequencing technologies with shorter reads.

1.1.2.2 Impact of Genomics

Genomics has stimulated investigations into many of these previously largely explored areas of functional genetics and thus been of immense value in genetics research, attracting high expectations for further applications. Genomics research has become anticipated for it's potential for widespread applications in healthcare, agriculture, ecology, conservation, and evolutionary biology, although many of these are yet to come to fruition.

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

1.1.3 Technologies to Enable Genetics Research

1.1.3.1 DNA Sequencing and Genotyping Technologies

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

The Sanger (dideoxy) chain termination method (Sanger and Coulson, 1975) enabled DNA sequencing and genotyping at a widespread scale, being less technically difficult than the Maxam-Gilbert sequencing by degradation method (Gilbert and Maxam, 1973; Maxam and Gilbert, 1977), which required more radioactive and toxic reactants. The Sanger methodology has relatively long read length (particularly compared to early versions of more recent technologies), with read lengths of 500–700 base pairs accurately sequenced in most applications, usually following targeted amplification with PCR. Sanger sequencing by gel electrophoresis takes around 6-8 hours and has been further refined with the "capillary" approach to 1–3 hours and requiring less input DNA and reactants. The capillary approach has been scaled up to run in parallel from a 96 well plate, at 166 kilobases per hour. The 96 well parallel capillary method was one of the main innovations which made the first Human Genome Project feasible and was used throughout (Lander *et al.*, 2001). Due to the quality of the Sanger sequence reads and low cost, it is still widely used in smaller scale applications, clinical testing, and to validate the findings of newer approaches.

1.1.3.2 Microarrays and Quantitative Technologies

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

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

In addition to being more versatile and higher-throughput than PCR based techniques, microarrays are considered cost-effective, particularly when scaled up to large number of probes. They are also available with established gene panels or customised probes from a number of commercial manufacturers. These remained popular during the introduction of newer technologies due to reliability and this relatively lower cost, especially in large-scale projects involving many samples. However, microarrays have issues with signal-to-noise ratio, with both sensitivity to low nucelic acid abundance and "saturation" of probes at high abundance, edge effects, and requiring more starting material than qPCR. Thus qPCR is still used for many small gene panel studies.

1.1.3.3 Massively Parallel "Next Generation" Sequencing

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

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

However, the methodology itself has challenges with the sample preparation, requiring a relatively high quantity of input material and "contamination" with over abundant ribosomal rRNA taking up the majority of the sequencing if not purified correctly. This abundance of rRNA is a particularly important issue in microarray and RNA experiments in Eukaryotes where it is commonplace target the mRNA by binding to the poly-A tail (RNA-Seq) or 5' cap (CAGE-Seq). However, this has the potential to exclude microRNAs (miRNA) and long non-coding RNAs (lncRNA) of interest unless

the sample is prepared specifically to study these. Similarly capturing a subsection of the genome for exome analysis or reduced representation bisulfite sequencing (RRBS), focuses on sequencing DNA sequences and methylation levels of CpG sites near known genes to reduce cost, noise, and incidental findings.

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

1.1.3.3.1 Molecular Profiling with Genomics Technology

NGS is highly adaptable to different applications: DNA sequencing (whole genome or exome), DNA methylation (bisulfite-Seq), RNA-Seq, miRNAs, lncRNA, or chromatin immunopreciptation (CHIP-Seq). Employing RNA-Seq to the transcriptome are a common adaptation, RNAs is reverse transcribed and sequenced from the resulting complementary "cDNA". This is ulitised to be quantify the levels of RNA and identify which regions of DNA are expressed. Similar bisulfite treatment converts cytosine residues to uracil (sequenced as thymidine), sparing methylated cytosine enabling it to be distinguished with bisulfite-Seq for high-throughput detection of the notable epigenetic mark and is a common procedure to generate an epigenome. Subsets of the nucelic acid may be extracted for sequencing such the coding regions of DNA (for the "exome"), the mRNA 5́cap (CAGE-Seq), mRNA 3́poly-A tail (RNA-Seq), microRNA, or an enriched subset of variable regions for DNA sequencing ("genotyping by sequencing") and methylation studies ("reduced-representation bisulfite sequencing). High-throughput gel and mass spectrometry techniques have been employed to pro-

teins and metabolites to generate the proteome and metabolome respectively. These "omics" technologies are applicable across a wide range of biomolecules in a cell and these "molecular profiles" are produced in many experimental laboratories.

1.1.3.3.2 Established Sequencing Technologies

454 sequencing (acquired by Roche) commercially released from 2005 to 2013 was the first NGS technology, generating a vast 1 million reads per day or 400–600Mbp in a 10 hour run. This technology used the "pyrosequencing" method of sequencing by synthesis, detecting phosphates released when a compatible nucleotide reacts and extends the DNA synthesis of a complementary strand. This technology popularised NGS with the first complete genome from a single individual (Wadman and Watson, 2008; Wheeler *et al.*, 2008) and Neanderthal ancient DNA studies (Green *et al.*, 2009; Noonan *et al.*, 2006). While this technology was capable of reads up to 1kb, reads of 400–500bp were more typical and the technology had difficulties with accurately processing runs of repeated bases (Rothberg and Leamon, 2008). These are still relatively long reads for an NGS technology but it has been discontinued due to competing short read technologies being more cost-effective with lower running costs.

SOLiD sequencing (acquired by Life Technologies and then Thermo Fisher) released in 2006 employed a vastly different approach to NGS, using labelled dinucleotide pairs for "sequencing by ligation" to produce a highly accurate sequence (99.94%) with built-in error correction by sequencing two reading frames and is unaffected by consecutive bases. This technology is also high-throughput, producing 1200–1400 million reads (66–120Gbp) in a 7–14 day run (ThermoFisher, 2017b). However, SOLiD sequencing does not cope well with palindromic sequences and SOLiD reads are very short only 35bp, making it more difficult to assemble them.

Illumina sequencing (developed by Solexa and later acquired by Illumina) was also released in 2006. It utilises reversible terminating dyes to sequence by synthesis with a lower accuracy (98%) and read lengths of 150–250bp. Illumina more than makes up for relatively short reads (along with improving the read length of the technology) and low accuracy with high-throughput and cost effectiveness, with a Hi-Seq 4000 platform producing up to 10 billion paired-end reads (1500Gbp) in a run of appropriately 3 days, capable of sequencing 12 human genomes (30× coverage) or 100 human transcriptomes simultaneously (Illumina, 2017). Illumina has further reduced the cost of sequencing with the economies of scale with the Hi-Seq X 10 claiming to produce a human genome (with 30× coverage) for less than US$1000, the first platform to achieve this long-

standing goal in genomics. The high-throughput of Illumina sequencing also makes deep sequencing for high coverage, high quality consensus reads, and sensitive RNA-Seq experiments feasible. Illumina sequencing now has a dominating market share of the NGS technologies.

1.1.3.3.3 Emerging Sequencing Technologies

Ion Torrent (also acquired by Life Technologies) released in 2010 employs "sequencing by synthesis" but in a drastically different way with ion semiconductor sequencing, detecting H^+ ions released when bases during DNA synthesis. Without the use of optical detection, the Ion Torrent system is compact offering rapid, cost-effective sequencing with the potential to scale with the future development of silicon semiconductors with have historically doubled in density every 2 years (Moore's Law). It is capable of reads of 100–200bp in only an hour (as fast as 4 seconds per base) and up to 400bp in a 2 hour run with an accuracy of 99.6% (dropping to 98% for consecutive sequences of 5 bases). While fast, cost effective, and accurate, Ion Torrent has short reads and modest throughput (up to 10 Gp for the Ion Proton and 15 Gb for the Ion S5 XL systems) compared to other sequencing technologies (ThermoFisher, 2017a).

Pacific Biosciences (PacBio) released the RS and RS II platforms in 2010 and 2011 to make up for the short reads in NGS technologies with the single molecule real time (SMRT) approach capable of long read lengths, averaging between 2.5–7kb and up to 80kb PacBio (2017). The PacBio methology traps each molecule in a zero mode waveguide (ZMW) and sequences it in real time. The RS II has 150,000 ZMW and an output of 500Mbp–1Gbp per SMRT cell (doubling that of the RS), with the capacity to run up to 16 concurrently for 0.5–6 hours. While the single molecule sequencing approach has strengths in sensitivity and potential to detect 3D structures, such as G-quadruplexes, this has the drawback of slowing down the sequencing and reducing the throughput of the platform. Another issue is sequence quality with the raw data as poor as 20–30%. However, PacBio recommends specific software to assemble as consensus with 99.999% for sequences with over 20× coverage, regardless of sequence repeats or GC composition. Despite concerns over data quality and higher cost than other approaches, the long reads are appealing for genome assembly and in many genome studies combine PacBio reads with more accurate short read technologies. However, due to the poor separate quality of reads this technology may not be appropriate for RNA-Seq studies, while it does have the potential for high sensitivity and detecting alternative splicing were it be improved. PacBio has recently released the Sequel (2016)

system, increasing the throughput of the SMRT Cells 7× to 1 million ZMW holes with an output of 5–10Gb for each of 16 SMRT cells.

Nanopore sequencing is another technology capable of long reads in real time and direct single molecule sequencing, avoiding amplification bias, detecting modified bases and directly sequencing RNA molecules. This also reduces laboratory preparation times. Nanopores work by measuring the ion current through a pore in a electrically insulating membrane as a nucleic moves through it. Oxford Nanopore has been developing this technology since 2005, launching the MinION in 2014 which employs biological nanopores: a transmembrane protein through which DNA or RNA passes, blocking ion current differently for each base. Each pore sequences in real time, capable of sequencing 450bp per second (Nanopore, 2017). However, the are quality issues with each individual read with quality estimates varying between 87–98%, with improvements to the quality of detection accounting for significant delays in the release of this technology. The MinION makes up for the is a capacity for extremely long reads, averaging 5.4kbp (Hayden, 2014) up to a maximum of 200Kbp and being a portable platform with very few overhead costs. While the MinION is limited in scale with only one flow cell of 512 pores (5–10Gbp), the PromethION being released in early access in 2016 scales this technology with flow cells of 3000 pores and the capacity to run 48 (up to 4 samples each) in parallel for 144,000 long reads with a versatile, modular system including built-in computing resources. One of the main issues with Oxford Nanopore systems is accuracy, with the manufacturer suggesting the use of consensus sequences for higher accuracy as PacBio does. The main source of this pore accuracy is the width of biological pores resulting in several bases being in the pore at any one time, inferring the sequence from the ion currents of each respective combination of bases and distinguishing them is a major technical challenge.

Quantum Biosystems in Japan is developing a synthetic nanopore system to address this issue. While the technology is still in development, it has the potential to produce similarly long reads, with a high-throughput, low running cost, and rapid run time (Quantum Biosystems, 2017). The technical challenges to develop a nanotechnology capable of this are immense but such developments serve as but one of example of how sequencing technologies may continue to improve, becoming more feasible for a wider variety of applications.

Due to such benefits of sequencing over previous technologies (and their continued refinement), this thesis has focused on gene expresseion data generated by RNA-Seq rather by microarrays. RNA-Seq data is widely available as a resource from large-scale

cancer genomics projects and methods to make inferences from RNA-Seq experiments could feasibly be applied to many other studies based on these current (or similar future) technologies.

1.1.3.4 Bioinformatics as Interdisciplinary Genomic Analysis

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

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

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

1.1.4 Follow-up Large-Scale Genomics Projects

A number of projects have attempted to follow up on the human genome project to varying degrees of success. The genomes have since been sequenced for a variety of model organisms, organisms of importance in health, agriculture, metagenomics of microorganisms (microbiome), ecology and conservation. Genomics projects have also

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

Other genomics databases have focused on faciltating distribution of genomic data generated by researchers, rather than generating it themselves. Genbank (NCBI) in the US, EMBL in Europe, and the DDBJ (NIG) in Japan do so by serving as repositories of DNA sequence data. GEO (Clough and Barrett, 2016), arrayExpress (Rustici *et al.*, 2013), and caArray (Heiskanen *et al.*, 2014) serve a similar purpose as a resource for gene expression datasets, originally developed for microarray data but RNA-Seq data is now supported by some platforms. They are repositories for researchers to deposit, share, and access gene expression data, which serve as a resource to support ongoing research to ulitise data for genes of interest to particular research groups and further to make inferences based on larger datasets than accessible to any individual laboratory (Rung and Brazma, 2013). These resources cover not only DNA sequence across the genome but also molecular profiles of other factors by adapting genomic sequencing or other high throughput technologies for quantifying gene expression or DNA methylation. Sharing the expression datasets generated in a publication is now required by some journals.

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

1.1.5 Cancer Genomes

It's importance in the future of cancer research was noticed, even in the early days of genomics (Dickson, 1999). The Cancer Genome Project (CGP) based at Wellcome Trust Sanger Institute in the UK were among the first to launch investigations into cancer after the publication of the Human Genome, using this genome sequence, consensus across the cancer reserach literature, and sequencing the genes of cancers themselves.

Initially, the Sanger Institute set out to sequence 20 genes across 378 samples while the Human Genome project was still ongoing (Collins and Barker, 2007), optimising sequencing and computation infrastructure for a larger project while doing so. The main aim of the Cancer Genome Project was to discover "cancer genes", those frequently mutated in cancers by comparing the genes of cancer and normal tissue samples, both "oncogenes" and "tumour suppressors" which are activated and inactivated respectively in cancers. This project is ongoing and the UK continues to be involved in international sequencing initiatives and those focused on particular tissue types.

The Sanger Institute also hosts the Catalogue of Somatic Mutations in Cancer (COSMIC, 2016), a database and website of cancer genes. This launched with 66,634 samples and 10,647 mutations from initial investigations into *BRAF*, HRAS, *KRAS*2, and NRAS (Bamford *et al.*, 2004). It has since expanded to include 1,257,487 samples with 4,175,8787 gene mutations curated from 23,870 publications, including 29,112 whole genomes (COSMIC, 2016). This database now also identifies cancer genes from DNA copy number, differential gene expression and differential DNA methylation.

1.1.5.1 The Cancer Genome Atlas Project

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

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

1.1.5.2 The International Cancer Genome Consortium

TCGA and the Cancer Genome project in the UK are part of a larger International Cancer Genome Consortium (ICGC), now a concerted effort across 16 countries to sequence the genome, transcriptome, and epigenome of 50 tumour types from over 25,000 samples total (Zhang *et al.*, 2011). With some redundancy the following countries are profiling various tumour types: USA (including TCGA), China (16), France (10), Australia (4), South Korea (4), the UK (4), Germany (4), Canada (3), Japan (3), Mexico (3 in collaboration with the US), Singapore (2), Brazil, India, Italy, Saudi Arabia, and Spain. This is inherently international and several projects are collaborations, such as between the USA and Mexico, Australia and Canada, Singapore and Japan, along with the UK and France representing the European Union (ICGC, 2017). In order to avoid competing the existing TCGA projects, some countries focus on a particular cancer they have health interest: Australia (melanoma), Brazil (melanoma), India (oral), Saudi Arabia (thyroid), and Spain (CML). Others focus on a particular tissue subtype with poor prognosis: The UK (triple negative or Her2+ breast cancer), France (clear cell kidney), Australia and Canada (ductal Pancreas). Another approach is to focus on rare or child cancers: Canada, Italy, France, Germany, Japan and Singapore, and the US (TARGET project). Particularly countries in Asia (China, Japan, Singapore, and South Korea) have emphasised the value of adding tumour data from non-Western countries or non-European populations in addition the data from Europe and the TCGA in the US. Data from 9 of these countries is already available on the ICGC website with the project ongoing.

1.1.5.2.1 Findings from Cancer Genomes

The cancer genome atlas pilot projects (Bell *et al.*, 2011; Hammerman *et al.*, 2012; McLendon *et al.*, 2008) serve to demonstrate the power of applying genomics technologies to cancer research at such as scale. In addition to sequence the whole genome or a subset (exome), DNA copy number, gene expression, DNA methylation, and somatic mutations were also analysed. The initial projects used microarray technologies for expression and methylation data but these have since been replaced by RNA-Seq for expression. TCGA demonstrated the potential discovery of the molecular basis of cancer by analysing 206 glioblastoma brain cancer samples (McLendon *et al.*, 2008), highlighting the roles of *ERBB2*, *NF1*, *TP53*, and *PIK3R1* mutations, along with altered methylation of *MGMT*, and the core pathways of RTK, p53, and RB signaling

in brain cancer. An analysis of 489 serious ovarian cancers (Bell *et al.*, 2011) similarly reported *TP53* mutations specifically over-represented in high grade tumours and reported 133 copy number variants, 168 differentially methylated regions, and recurrently somatic mutations in 9 genes in low grade tumours including *NF1*, *BRCA1*, *BRCA1*, *RB1*, and *CDK12*. Four transcriptional subtypes of ovarian cancers were identified, alterations in *BRCA1*, *BRCA2*, and *CCLE* had an impact of patient survival, and the homologous recombination, NOTCH and FOXM1 signaling pathways were involved in ovarian cancer growth. The genomics of 178 squamous cell lung cancers (Hammerman *et al.*, 2012) were highly complex, averaging at 360 muations in coding regions. While no targeted therapies existed for this cancer subtype, 11 recurrently mutated genes were identified including *TP53* and *HLA-A*. The pathways altered in various squamous cell lung cancers were NFE2L2, KEAP1, differentiation genes, PI3K, CDKN2A and RB1. These abberant genes and pathways represent potential therapeutic targets which could be identified for most samples.

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

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

The TCGA stomach cancer analysis of 295 samples (Bass *et al.*, 2014) identified 4

molecular subtypes of stomach cancers characterised by: the Epstein-Barr virus, MSI, genomics instability, and chromosomal instability. Abberrations in *PD-L1*, *PIK3CA*, and *JAK2* were also identified in stomach cancers which may present therapeutic targets.

1.1.5.2.2 Genomic Comparisons Across Cancer Tissues

TCGA have identified various genes as recurrent, driver mutations across cancer types which are likely to have a role in driving the proliferation of these cancers and present a molecular target that could be applied across tissue types. These include *TP53* (in brain, lung/head/neck squamous cell, breast, colorectal, uterine, and endometrial cancers), *ERBB2*/HER2/NEU (in brain, breast, colorectal, bladder, and lung cancers), *PIK3CA, PIK3R1* (in brain, breast, colorectal, endometrial, bladder, clear cell renal, and lung cancers), *BRCA1*/*BRCA2* (in breast and ovarian cancers), *NF1* (in brain, ovarian, and skin cancers), *ARID1A* (in colorectal, endometrial, and clear cell renal cancers), *KRAS* (in colorectal, endometrial, and skin cancers), *BRAF* (in colorectal, thyroid, and skin cancers), *EGFR* (in brain, breast, and lung cancers), and *PTEN* (in breast, endometrial, and uterine cancers) (Agrawal *et al.*, 2014; Akbani *et al.*, 2015; Bass *et al.*, 2014; Bell *et al.*, 2011; Burk *et al.*, 2017; Cherniack *et al.*, 2017; Ciriello *et al.*, 2015; Collisson *et al.*, 2014; Creighton *et al.*, 2013; Hammerman *et al.*, 2012; Kandoth *et al.*, 2013; Lawrence *et al.*, 2015; McLendon *et al.*, 2008; Muzny *et al.*, 2012; TCGA, 2012; Weinstein *et al.*, 2014). In addition to disregarding the distinct between colon and rectal cancers based on molecular similarlity (Muzny *et al.*, 2012), the TCGA project have observed differences within tumour types and proposed molecular subtyping for breast, clear cell renal, papillary renal, stomach, skin, bladder, and prostate cancers (Abeshouse *et al.*, 2015; Akbani *et al.*, 2015; Bass *et al.*, 2014; Ciriello *et al.*, 2015; Creighton *et al.*, 2013; Hammerman *et al.*, 2012; Linehan *et al.*, 2016; Muzny *et al.*, 2012; TCGA, 2012; Weinstein *et al.*, 2014).

The "Pan cancer" project (Hoadley *et al.*, 2014; Weinstein *et al.*, 2013) analysed 3527 samples across 12 tissue types for DNA, RNA, protein, and epigenetic molecular profiles. This project was initated in 2012 to perform a comprehensive analysis of molecular data across cancer types to identify molecular simliarities and differences. Recurrent *TP53* mutations characterised high grade tumours across breast, ovarian, and endometrial cancers. HER2 was identified in brain, endometrial, bladder, and lung cancers, in addition to the known role of HER2 in breast cancers. *BRCA1* and *BRCA2* mutations were also detected across cancers, mainly breast and ovarian cancers

as expected. Microsatellite instability characterised both endometrial and colorectal cancers. The Pan cancer project (Hoadley *et al.*, 2014) has identified 11 molecular subtypes across these tissues, 5 of corresponding to tissue cancer types and the remainder reassigned due to molecular similarities shared across cancer types. Squamous cell lung, head, and neck and a subset bladder cancers were grouped together by molecular similarities, characteried by a high frequency of *TP53* mutations. Conversely, bladder cancers were divided into 3 of these molecular subtypes with distinct profiles. This project further supports the genomic stratification of patients, demonstrated in breast cancer (Parker *et al.*, 2009; Pereira *et al.*, 2016; Perou *et al.*, 2000), which may apply to other cancer types and to molecular characteristics across them targeting recurrent mechanisms of cancer growth and progression (Hanahan and Weinberg, 2000, 2011).

1.1.5.2.3 Cancer Genomic Data Resouces

While the findings from the TCGA projects themselves are a considerable contribution to understanding cancer biology within and across tissue types, the main eventual benefit of such projects will be the avaialability of the data for the research community to analyse further and use to inform future investigations (McLendon *et al.*, 2008; TCGA, 2017; Weinstein *et al.*, 2013). These serve as a vast resource of common and rare cancer types and are publicly available to analyse further (cBioPortal, 2017; TCGA, 2017; Zhang *et al.*, 2011). This also applies to the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project which focuses on breast cancer which also aimed to identify novel molecular subtypes (Curtis *et al.*, 2012). They performed an analysis of 2433 breast cancer samples with long-term clincal data, gene expression, copy number variants, and 173 genes sequenced which identified 40 driver mutations in breast cancer in addition to further support for molecular subtyping to identify patient groups with different clinical outcomes (Pereira *et al.*, 2016).

1.1.6 Genomic Cancer Medicine

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

in public healthcare systems with a limited budget.

1.1.6.1 Cancer Genes and Driver Mutations

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

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

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

1.1.6.2 Personalised or Precision Cancer Medicine

The notion of using a patient's genome to tailor healthcare to an individual has been appealing since the advent of genomics, popularised with the term "personalised medicine". This approach was expected to span from preventative lifestyle advice to effective treaments. Personalised medicine was intended contrast with current strategies of health advice, screening, prognostics, and treaments based on what works well with the majority of the population, highlighting that adverse effects of treaments occur in a significant subpopulation and that many clinical studies are dominated by Western populations of European ancestry and may not generalise to other populations.

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

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

1.1.6.2.1 Molecular Diagnostics and Pan-Cancer Medicine

There is growing support for the use of molecular tools such as mutations or gene expression signatures to diagnose tumour subtypes in replacement or addition to tissue of origin or histology. This is particularly important in breast cancer where analysis of

molecular data detected several distinct "intrinsic subtypes" with differences in malignancy and patient outcome which were distinguished by molecular mechanisms rather than tissue or cellular phenotype (Parker *et al.*, 2009; Perou *et al.*, 2000). Conversely, common molecular mechanisms may be shared between cancers across tissue types as discovered by the "Pan cancer" studies, such as those conducted by the TCGA and ICGC projects, which combined molecular profiles across tissue types Weinstein *et al.* (2013). The molecular subtypes could feasibly be included in clinic testing as a panel of biomarkers for diagnostics and prognosis. Such biomarkers also have the potential to monitor drug response or risk of recurrence. This is also raises the need for development of treatments for targeting these molecular subtypes.

1.1.6.3 Targeted Therapeutics and Pharmacogenomics

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

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

1.1.6.3.1 Targeting Oncogenic Driver Mutations

Oncogene targeted therapies have also been developed with some examples of effective clinical application against cancers. However, they already begun to manifest problems with resistance, recurrence, tissue specificity, and design of inhibitors specific to oncogenic variants rather than proto-oncogene precusors. Targeted anticancer therapeutics can exploit complex interactions to distinguish normal and cancerous cells which may benefit from studies of gene regulation or interaction networks. The unexpected syn-

ergy between inhibitors of the oncogenes $BRAF^{V600E}$ and *EGFR* in colorectal cancer is an example of such a system Prahallad *et al.* (2012).

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

1.1.6.4 Systems and Network Biology

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

The properties of large networks were first studied by constructing random networks by randomly linking a fixed number of nodes (Erdős and Rényi, 1959, 1960). Despite the random nature of these networks, properties such as their connectivity were well characterised. The vertex degree (number of partners for each node) of random network follows a Poisson distribution, however this property does not hold in nature, suggesting that natural networks are non-random or not formed in this way Barabási and Oltvai (2004).

This work formed the foundation for studying complex networks (van Steen, 2010), which model features of real world networks not found in Erdős and Rényi‘s random networks (Erdős and Rényi, 1959, 1960). The small world property, made popular by findings in social networks (Travers and Milgram, 1969), is the remarkably short path lengths between any nodes in a small world network. A small world network is

well-connected with a characteristic path length (the average length of shortest paths between all pairs of nodes) proportional to the logarithm of the number of nodes. Watts and Strogatz (1998) developed a model of random rewiring of a regular network to construct random networks with the small world property and a high clustering coefficient. While these properties are more representative of networks occurring in nature, their model is limited by the degree distribution which converges to a Poisson distribution as it is rewired Barrat and Weigt (2000).

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

Barabási and Albert (1999) constructed a network model in an entirely different way to randomly generate scale-free networks which have a power law degree distribution. They constructed random networks by preferential attachment, modelling growth of a network by sequentially adding nodes with links to existing nodes. The scale-free nature of the random networks was ensured by adding new nodes with an increasing probability of attachment to an existing node if it has higher degree. These networks successfully capture the scale-free nature of many real world networks with short characteristic path length and low eccentricity resulting in super small worlds Barabási and Albert (1999). Scale-free networks are limited by a low clustering coefficient and lack of modular structure; however, they have enabled the study of scale-free network topology and served as a basis for modified scale-free models (Dorogovtsev and Mendes, 2003; Holme and Kim, 2002).

Han *et al.* (2004) observed dynamic modularity in biological networks and suggested the network structure may underpin genetic robustness and plasticity. They focus on network hubs which are more likely to be essential genes and define the subgroups of hubs based on correlation of gene expression with protein-protein interaction partners: 'party' hubs (which interact simultaneously with many partners) and 'date' hubs (which interact with different partners in different conditions). Party and date hubs occurred

most frequently within and between network modules respectively. Party hubs were considered local regulators, whereas date hubs were considered important to network connectivity as global regulators. This distinction between classes of network hubs was supported by differences in tissue specificity and clinical relevance as a proposed predictor of clinical outcome in breast cancer with an AUROC of 0.784 Taylor *et al.* (2009). However, correlation between expression and protein interactions were not robustly reproduced. The importance of date hubs has been criticised for assuming a bimodal distribution and basing the global importance of data hubs on a small subset Agarwal *et al.* (2010). As an alternative interpretation, (Agarwal *et al.*, 2010) suggest the importance of interactions rather than network hubs as interactions important to the network were between functionally similar proteins. Network hubs can also be classed as associative or dissociative depending on whether they tend toward or away from connecting directly to other network hubs (van Steen 2010). The associative and dissociative properties can also be used to test whether nodes of a particular subgroup (e.g., gene function) associate with each other.

Applications of network theory are diverse, including uses in social sciences, engineering, and computer science. Due to their complexity and difficulty of gathering sufficient empirical data, biological applications of network theory are relatively unexplored. High-throughput technologies such as siRNA screens, two-hybrid screens, microarrays and massively parallel sequencing have made generating genome-scale molecular data feasible and enabled analysis of biological networks at the molecular level. Many types of inter-related molecular networks can be constructed and analysed, depending on the biological application. Genetic interaction networks will be the focus of this project because they are relatively unexplored compared to other molecular networks, have potential for applications in drug discovery (particularly cancer treatment), and may lead to better understanding of the role of genetics in cellular function and disease. Genetic interactions are usually studied at a high-throughput scale in simple model organisms such as bacteria, yeasts or the nematode worm; studies in humans, mammals, and non-model organisms (where applications would have the most societal impact) are limited by cost, time and labour constraints. Computational approaches with effective predictive models are the only feasible approach to study the connectivity of a biological network in a complex metazoan cell at the genome-scale.

1.1.6.4.1 Network Medicine, and Polypharmacology

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

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

A molecular map of the interactions and pathways in the mammalian cellular network has the potential to impact upon drug design and clinical practice, particularly in treatment of cancer and infectious disease. Characterisation of the target system and impact of existing treatments, such as $BRAF^{\mathrm{V600E}}$ and *EGFR* inhibitors, enable wider application of the mechanisms for such interventions exploiting genetic interactions or pathways. This could lead to development of more effective treatment interventions for these systems and prediction of similar molecular systems for development of novel drug targets and combinations.

1.2 A Synthetic Lethal Approach to Cancer Medicine

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

1.2.1 Synthetic Lethal Genetic Interactions

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

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

viability, "suppression" and "rescue" are positive SGIs giving at least partial restoration of wild-type growth from single mutants with growth impairment and lethal phenotypes respectively. Negative SGIs were markedly more common than positive SGIs in a number of studies in model systems Boucher and Jenna (2013); Tong *et al.* (2004).

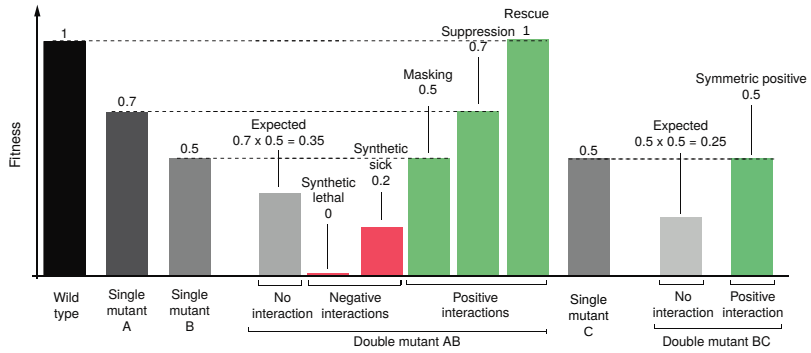


Figure 1.1: **Synthetic genetic interactions.** Impact of various negative and positive SGIs: negative interactions involve deleterious (sick) or inviable (lethal) phenotypess whereas positive interactions involve restoring viability by masking or suppressing the other mutation or complete rescue of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

1.2.2 Synthetic Lethal Concepts in Genetics

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

Since synthetic lethal gene partners can be affected by extracellular stimuli and chemical, essentiality of synthetic lethal genes can be induced by the environment of a cell. An environmental stress conditions may inhibit one or the other synthetic lethal gene, such as exposure to chemicals, in which case the synthetic lethal partner gene is "conditionally essential" (Hillenmeyer, 2008). Thus the evolutionary rationale for the abundance of SGIs (compared to the surprisingly low number of essential genes) in a

Eukaryotic genome attributed to genetic functional redundancy and network robustness of a cell which are advantageous to survival.

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

1.2.3 Studies of Synthetic Lethality

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

1.2.3.1 Synthetic Lethal Pathways and Networks

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

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

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

1.2.3.1.1 Conservation and Evolution of Synthetic Lethality

There is poor conservation of specific SGIs between *S. cerevisiae* and *S. pombe* with 29% of the interactions tested in both distantly related species being conserved between them (Dixon2008). The remaining interactions show high species-specific differences; however, many of the species specific interactions were still conserved between biological pathways, protein complexes, or protein-protein interaction modules. Similarly, conservation of pathway redundancy was also found between Eukaryotes (*S. cerevisiae*) and prokaryotes (*E. coli*) (Butland *et al.*, 2008). Negative SGIs were more likely to be conserved between biological pathways, whereas positive SGIs were more likely to be conserved within a pathway or protein complex (Roguev *et al.*, 2008).

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

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

1.2.4 Synthetic Lethal Concepts in Cancer

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

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

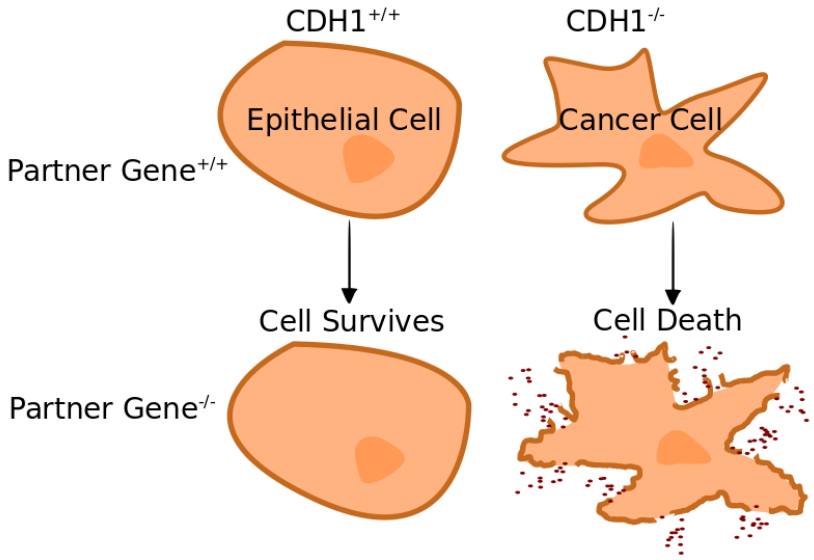


Figure 1.2: **Synthetic lethality in cancer.** Rationale of exploiting synthetic lethality for specifity against a tumour suppressor gene (e.g., *CDH1*) while other cells are spared under the inhibition of a partner gene.

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

1.2.5 Clinical Impact of Synthetic Lethality in Cancer

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

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

somatic mutations in some familial and sporadic cancers. Poly-ADP-ribose polymerase (PARP) genes are tumour suppressor genes involved in base excision DNA repair. Loss of PARP activity results in single-stranded DNA breaks. However, *PARP1*$^{-/-}$ knock-out mice are viable and healthy indicating low toxicity from PARP inhibition (Bryant *et al.*, 2005).

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

Thus PARP inhibitors have potential for clinical use against *BRCA1* or *BRCA2* mutations in hereditary and sporadic cancers (Ashworth 2008; Kaelin2005). PARP inhibition has been found to be effective in cancer patients carrying *BRCA1* or *BRCA2* mutations and some other ovarian cancers, suggesting synthetic lethality between PARP and other DNA repair pathways (Ström and Helleday, 2012). This supports the potential for PARP inhibition as a chemo-preventative alternative to prophylactic surgery for high risk individuals with *BRCA1* or *BRCA2* mutations (Ström and Helleday, 2012). Hormone-based therapy has also been suggested as a chemo-preventative in such high risk individuals and aromatase inhibitors have completed phase I clinical trials for this purpose (Bozovic-Spasojevic2012). Ström and Helleday (2012) also postulate increased efficacy of PARP inhibitors in the hypoxic DNA-damaging tumour micro-environment.

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

toxicity between carriers of *BRCA1* or *BRCA2* mutations and sporadic cases. AstraZeneca announced that olaparib has begun phase III trials for breast and ovarian cancers in 2013. Mixed results in phase II trials in ovarian cancer are behind the delays addressed by retrospective analysis of the cohort subgroup with confirmed mutation of *BRCA1* or *BRCA2* genes in the tumour; unsurprisingly these patients, benefit most from the PARP inhibitor treatment and have increased platinum sensitivity in combination treatment. This demonstrates the clinical impact of a well characterised system of synthetic lethality with known cancer risk genes. Synthetic lethality has the benefit of being effective against inactivation of tumour suppressor genes by any means, broader than targeting a particular oncogenic mutation (Kaelin, Jr, 2005). The targeted therapy is effective in both sporadic and hereditary *BRCA1* or *BRCA2* deficient tumours acting against an oncogenic molecular aberration across several tissues.

[Update re. FDA approval for Ovarian]

These PARP inhibitors are FDA approved for some cancers McLachlan *et al.* (2016), are effective against germline and sporadic *BRCA1* or *BRCA2* mutations, and are a potential preventation alternative to prophylactic surgery for high risk mutation carriers Ström and Helleday (2012).

1.2.6 High-throughput Screening for Synthetic Lethality

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

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

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

can be approached with genetic screens, so we will focus on these and computational alternatives. Screening methods have proven a fruitful area of research, despite being costly, laborious, and having many different sources of error. These limitations suggest a need for complementary computational approaches to synthetic lethal discovery.

1.2.6.1 Examples of Synthetic Lethal Screens

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

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

Similarly, hereditary diffuse gastric cancer (HDGC) is a cancer syndrome of predisposition to early-onset malignant stomach and breast cancers attributed to mutations in E-cadherin (*CDH1*). Telford *et al.* (2015) performed an RNAi screen on MCF10A breast cells for synthetic lethality with *CDH1*. They found enrichment of G-protein coupled receptors (GPCRs) and cytoskeletal gene functions. The results were consis-

tent with a concurrent drug compound screen with a number of candidates validated by lentiviral shRNA gene knockdown and drug testing including inhibitors of Janus kinase, histone deacetylases, phosphoinositide 3-kinase, aurora kinase, and tyrosine kinases. Therefore the synthetic lethal strategy has potential for clinical impact against HDGC, with particular interest in interventions with low adverse effects for chemoprevention, including repurposing existing approved drugs for activity against *CDH1* deficient cancers.

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

Chemical genetic screens are also a viable strategy to identify therapeutically relevant synthetic lethal interactions. Bitler *et al.* (2015) investigated *ARID1A* mutations, aberrations in chromatin remodelling known to be common in ovarian cancers, for drug response. Ovarian RMG1 cells were screened for drug response specific to *ARID1A* knockdown cells. They used *ARID1A* gene knockdown for consistent genetic background, with control experiments and 3D cell culture to ensure relevance to drug

activity in the tumour micro-environment. Screening a panel of commercially available drugs targeting epigenetic regulators found *ESH2* methyltransferase inhibitors effective and specific against *ARID1A* mutation with validation in a panel of ovarian cell lines. Synthetic lethality between *ARID1A* and *ESH2* was supported by decreases in H3K27Me3 epigenetic marks and markers of apoptosis in response to *ESH2* inhibitors. This was mechanistically supported with differential expression of *PIK3IP1* and association of both synthetic lethal genes with the *PIK3IP1* promoter identifying the PI3K-AKT signalling pathway as disrupted when both genes are inhibited. This successfully demonstrates the importance of synthetic lethality in epigenetic regulators, identifies a therapeutically relevant synthetic lethal interaction, and shows that chemical genetic screens could model drug response and combination therapy in cancer cells. However this approach is limited to finding synthetic lethal interactions between genes with known similar function, which may not be the most suitable for treatment. Further limiting experiments to genes with existing targeted drugs reduces the number of synthetic lethal interactions detected, assumes on their drug specificity to a particular target, and many of these drugs are not clinically available yet anyway as they are still in clinical trials for other diseases or are not supported by healthcare systems in many countries.

The examples above show that high-throughput screens are an effective approach to discover synthetic lethality in cancer with a wide range of applications. Screens are more comprehensive than hypothesis-driven candidate gene approaches and successfully find known and novel synthetic lethal interactions with potential for rapid clinical application. They have the power to test mode of action of drugs, find unexpected synthetic lethal interactions between pathways, or identify effective treatment strategies without needing a clear mechanism. However, synthetic lethal screens are costly, labour-intensive, error-prone, and biased towards genes with effective RNAi knockdown libraries. Limited genetic background, lethality to wild-type cell during gene knockdown, off-target effects, and difficultly replicating synthetic lethality across different cell lines, tissues, laboratories, or conditions stems from a high false positive rate and a lack of standardised thresholds to identify synthetic lethality in a high-throughput screen. Therefore there is a need for replication, validation, and alternative approaches to identify synthetic lethal candidates. Varied conditions between experimental screens and differences between RNAi or drug screens makes meta-analysis difficult. Thus genome-scale synthetic lethal experiments are not feasible, even in model organisms, so a computational approach would be more suitable for this task.

1.2.7 Computational Prediction of Synthetic Lethality

1.2.7.1 Bioinformatics approaches to gene interactions

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

Table 1.1: Methods for Predicting Genetic Interactions

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

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

Table 1.2: Methods for Predicting Synthetic Lethality in Cancer

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

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

1.2.7.2 Comparative genomics

A comparative genomics approach by Deshpande *et al.* (2013) used the results of well characterised high-throughput mutation screens in *S. cerevisiae* as candidates for synthetic lethality in humans (Baryshnikova *et al.*, 2010a; Costanzo *et al.*, 2010, 2011; Tong *et al.*, 2001, 2004). Yeast synthetic lethal partners were compared to human orthologues to find cancer relevant synthetic lethal candidate pairs with direct therapeutic potential. Proposed as a complementary approach to siRNA screens, approximately 24,000 of the 116,000 negative SGIs in yeast (Costanzo *et al.*, 2011) were matched to human orthologues, with over 500 involving a cancer gene (Futreal *et al.*, 2004). Under strict criteria of one-to-one orthologues, large effect size and significant interaction in

yeast data ($\epsilon < -0.2$, $p < 0.05$), 1,522 interactions were identified with 70 involving cancer genes. Of the 21 gene interactions tested with pairs of siRNA in IMR1 fibroblast cells, 6 exhibited synthetic lethal effects. The two strongest interactions (*SMARCB1* with *PSMA4* and *ASPSCR1* with *PSMC2*) were successfully validated in by protein analysis of human cells and replication with tetrad analysis for yeast orthologues.

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

Machine learning approaches have also been proposed for synthetic lethal discovery (Babyak, 2004; Lee and Marcotte, 2009). Due to concerns that these may be subject to overfitting or noise, Wu *et al.* (2014) developed a meta-analysis method (based on the machine learning methods in Table 1.3) for synthetic lethal gene pairs rele-

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

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

vant to developing selective drugs against human cancer, building upon their previous database (Li *et al.*, 2014). The used training data of 10,885 synthetic lethal interactions from yeast experiments of which 7347 occurred between the 5,504 non-essential genes. Their "metaSL" approach utilises genomic, proteomic and annotation data (including GO terms Ashburner *et al.* (2000), PPI, protein complexes, and biological pathway) with strong stastical performance in yeast data (AUROC of 0.871). The predicted orthologous synthetic lethal partners in human data were not experimentally validated but several would be relevant to cancer such as *EGFR* with *PRKCZ*. They note that computational approaches scale-up across the genome at lower cost than experimental screen and share their most supported interactions online. However, the method is not available for analysis of other genes studied by the cancer research community. While machine learning has great potential as a predictor, the results vary greatly depending on the predictive features selected and it is not clear which threshold should be used to report reliably detected genes. Syn-Lethality (Li *et al.*, 2014) and MetaSL (Wu *et al.*, 2014) demonstrate the value of computational approaches to synthetic lethality but omit many genes of importance in cancer, such as *CDH1*, and there remains a need to enable biological researchers to query such genes in a particular tissue or genetic background.

There is also concern for analyses based on yeast data that many synthetic lethal

interactions may not be conserved between species Dixon *et al.* (2009), although interactions between pathways may are more comparable. It is unsurprising that many of the interactions identified were not experimentally validated. There have been many gene duplications in the separate evolutionary histories of humans and yeast which may lead to differences in genetic redundancy. Yeast are further not an ideal human cancer model because they are do not have tissue specificity, multicellular gene regulation, or orthologues to a number of known cancer genes such as p53. Although these studies have tried to anticipate these issues with stringent criteria such as requiring one-to-one orthologues, there remains the possibility that changes in gene function may affect whether these are solely redundant such as if functions had coevolved without sequence homology. Many genes will also be excluded by lacking homologous gene in yeast, the corresponding experimental data, or having paralogues in either species. Thus conservation of yeast interactions is not an ideal strategy and analysis of human data directly for comparison with human experimental data will be the focus of this thesis.

1.2.7.3 Analysis and modelling of protein data

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

However, their approach was limited to known cancer genes and is not applicable to genes that do not have PPI data. Other nucleotide sequencing data types are more commonly available for cancer studies at a genomic scale. Of further concern is that the results were enriched for p53 synthetic lethal partners which is relevant to many

cancer researchers but this genome-wide approach did not detect many other cancer genes due to multiple testing. This enrichment may be due to the known drastic effect of removing p53 itself from the network as a master regulator, cancer driving tumour suppressor gene, and highly connected network "hub". The focus on cancer genes is useful for translation into therapeutics but does not account for variable genetic backgrounds or effect of protein removal on the whole cellular network.

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

1.2.7.4 Differential gene expression

Differential gene expression has been explored to predict synthetic lethal pairs in cancer which would be widely applicable due to the availability of public gene expression data for a large number of samples and cancer types. Wang and Simon (2013) found differentially expressed genes (by the t-test, adjusted by FDR) between tumours with or without functional p53 mutations in TCGA (McLendon *et al.*, 2008) and Cell Line Encyclopaedia (CCLE) RNA-Seq gene expression data as candidate synthetic lethal partner pathways of p53. They identified 2, 8, and 21 candidate synthetic lethal partner genes in 3 microarray datasets from the NCI60 cell lines, 31 partner genes from the CCLE RNA-Seq data, and 50 in TCGA RNA-Seq data. *PLK1* was replicated across 4 of these analyses and 17 other genes were replicated across 2 analyses (including *MTOR*, *PLK4*, *MAST2*, *MAP3K4*, *AURKA*, *BUB1* and 6 CDK genes) with many playing a role in cell cycle regulation. This was supported by a drug sensitivity experiment on the NCI60 cell lines which found that cells which lacked functional p53 were more sensitive to paclitaxel (which targets *PLK1*, *AURKA*, and *BUB1*). This demonstrated the potential of gene expression as a surrogate for gene function and use of public genomic data to predict synthetic lethal gene pairs in cancer. Wang and Simon (2013) advocated for pre-screening of expression profiles to augment future RNAi screens. However, the analyses were limited to kinase genes and focused on currently druggable genes, lacking wider application of synthetic lethal prediction methodology. This approach may not be feasible or applicable in cancer genes with a lower mutation rate than p53.

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

Although the experimental work was the focus of the paper, these findings show that bioinformatics synthetic lethal candidates can be validated in patient tissue samples (from a non-caucasian population) to find those applicable to colorectal cancers.

1.2.7.5 Data mining and machine learning

Recognising the utility of synthetic lethality to drug inhibition and specificity of anti-cancer treatments, Jerby-Arnon *et al.* (2014) also saw the need for effective prediction of gene essentiality and synthetic lethality to augment experimental studies of SL. They developed a data-driven pipeline called DAISY (data mining synthetic lethality identification pipeline) and tested for genome-wide analysis of synthetic lethality in public cancer genomics data from TCGA and CCLE. DAISY is intended to predict the candidate synthetic lethal partners of a query gene such as genes recurrently mutated in cancer.

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

The DAISY methodology (Jerby-Arnon *et al.*, 2014) compares the results of analysis of several data types to predict synthetic lethality, namely: DNA copy number and somatic mutation for TCGA patient samples and CCLE cell lines. The cell lines were also analysed with gene expression and gene essentiality (shRNA screening) profiles. Genes were classed as inactivated by copy number deletion, somatic loss of function mutation, or low expression and tested for synthetic lethal gene partners which are ei-

ther essential in screens or not deleted with copy number variants. Co-expression is also used for synthetic lethality prediction based on studies in yeast (Costanzo *et al.*, 2010; Kelley and Ideker, 2005). Copy number, gene expression and, essentiality analyses are stringently compared by adjusting each for multiple tests with Bonferroni correction and only taking hits which occur in all analyses. This methodology was also adapted for synthetic dosage lethality by testing for partner genes where genes are overactive with high copy number or expression. As discussed above, the predictions performed well and an RNAi screen for the example of *VHL* in renal cancer validated predicted synthetic lethal partners of *VHL* demonstrating the feasibility of combining approaches to synthetic lethal discovery in cancer and using computational predictions to enable more efficient high-throughput screening. DAISY performs well statistically with a AUROC of 0.779 on a set of gene pairs with experimental screen data, although co-expression and shRNA functional examination contributes much less of this than the mutation and copy number analysis (AUROC 0.683 alone). However, this methodology is very stringent, missing potentially valuable synthetic lethal candidates, may not be applicable to genes of interest to other groups and the software for the procedure is not publicly released for replication.

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

Lu *et al.* (2015) critique the assumption of co-expression in the DAISY methodology and propose an alternative computational prediction of synthetic lethality based on machine learning methods and a cancer genome evolution hypothesis. Using DNA copy number and gene expression data from TCGA patient samples, a cancer genome evolution model assumes that synthetic lethal gene pairs behave in 2 distinct ways in response to an inactive synthetic lethal partner gene, either a "compensation" pattern where the other synthetic lethal partner is overactive or a "co-loss underrepresentation" pattern where the other synthetic lethal partner is less likely to be lost, since loss of

both genes would cause death of the cancer cell. During the cancer genome evolution as the cell becomes addicted to the remaining synthetic lethal partner due to induced gene essentiality. These patterns would explain why DAISY detects only a small number of synthetic lethal pairs, compared to the large number expected based on model organism studies (Boone *et al.*, 2007), and the disparity between screening and computationally predicted synthetic lethal candidates due to testing different classes of synthetic lethal gene pairs.

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

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

1.2.7.6 Bimodality

Wappett *et al.* (2016) demonstrate a multi-omic approach to identification of synthetic lethality in cancer with a strategy to detect bimodal patterns in molcular profiles. The release this solution as the Bimodal Subsetting Expression (BiSEp) R package Wappett (2014) which aims to detect subtle bimodal and non-normal patterns in expression data. Since loss of gene function is not consistently genetic, Wappett *et al.* (2016) advocate the use of gene expression (loss of mRNA) and deletion (loss of copy number) data in addition to mutation. The BiSEp procedure was demonstrate on an analysis of 881 cell lines from CCLE (Barretina *et al.*, 2012), 442 cell lines from COSMIC (Forbes *et al.*, 2015), and RSEM normalised RNA-Seq data for 178 TCGA lung patient samples (Collisson *et al.*, 2014). BiSEp was demonstrated to have significant enrichment of validated tumour suppressor, synthetic lethal gene pairs (detecting 76 experimentally supported gene pairs) and was improved (detecting 420) with expression data rather than relying on detecting loss of gene function by mutation or deletion. They identified interactions with genes relevant to cancer with support in experimental screens including *ERCC4* with *XRCC1*, *BRCA1* with *PARP3*, and *SMARCA1* with *SMARCA4*.

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

1.2.7.7 Rationale for further development

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

cers. As such, this thesis presents a query-based method, amenable to identification of candidate partners for a selected gene of functional or translational importance such as *CDH1*.

1.3 E-cadherin as a Synthetic Lethal Target

E-cadherin is a transmembrane protein (encoded by *CDH1*) with several characterised functions in the cytoskeleton and cell-to-cell signaling. Here we outline the key known functions of E-cadherin and it's importance in cancer biology. *CDH1* is a tumour supressor gene, with loss of function occuring in both familial (germline mutations) and sporadic (somatic mutations) cancers. As such, *CDH1* inactivation is a prime example of a genetic event that could be targeted by synthetic lethality for anti-cancer treatments. Most notably this includes patients at risk of developing hereditary breast and stomach cancers for which conventional surgical or cytotoxic chemotherapy is not ideal (due to impact of quality of life) and who have a known genetic aberration in their familial syndromic cancers. Effective treatments against *CDH1* inactivation would also benefit patients with sporadic diffuse gastric cancers since they often present with symptoms at a late stage.

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

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

1.3.1.1 Cytoskeleton

The primary function of *CDH1* is cell-cell adhesion forming the adherens junction, maintaining the cytoskeleton and mediating molecular signals between cells. The function of the adherens complex is particularly important for cell structure and regulation because it interacts with cytoskeletal actins and microtubules. The cytoskeletal role of E-cadherin maintains healthy cellular viability and growth in epithelial tissues in-

cluding cellular polarity. E-cadherin is not essential to cellular viability but loss in epithelial cells does lead to defects in cytoskeletal structure and proliferation. In addition to a central role in the adherens complex, E-cadherin is involved in many other cellular functions and thus *CDH1* is regarded as a highly pleiotropic gene.

1.3.1.2 Extracellular and Tumour Micro-Environment

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

1.3.1.3 Cell-Cell Adhesion and Signalling

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

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

E-cadherin has key roles in maintaining cellular structure and regulating growth, consistent with *CDH1* being a tumour suppressor gene. Loss of *CDH1* in epithelial tissues leads to disrupted cell polarity, differentiation, and migration. E-cadherin loss has

been identified as a recurrent driver tumour suppressor mutation in sporadic cancers of many tissues including breast, stomach, lung, colon, and pancreas tissue.

1.3.2.1 Breast Cancers and Invasion

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

1.3.3 Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer

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

HDGC affects less than 1 in a million people globally (Ferlay *et al.*, 2015) and

less than 1% of gastric cancers. However, HDGC is documented to affect several hundred families globally. E-cadherin mutations in the germline is implicated in 1-3% of gastric cancers presenting with a family history, varing between high and low incidence populations. E-cadherin is also mutated in 13% of sporadic gastric cancers.

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

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

1.3.4 Somatic Mutations

1.3.4.1 Mutation Rate

Estimates for the prevalence of *CDH1* somatic mutations in sporadic cancers varies. The Cancer Gene Census (Futreal *et al.*, 2004; Pleasance *et al.*, 2010) detected 994 distinct mutations in 10,143 tumour samples (at a rate of 7.52%), COSMIC (2016) detected 632 distinct mutations in 43,865 tumour samples (at a rate of 1.71%), and the NCI60 detected mutations in 13.2% of 53 cancer cell lines. While there is no consensus

on the prevalence of *CDH1* mutations, the vast variability of mutations is consistent with it's role as a tumour supressor and it has been found to be recurrently mutated in a wide range of cancers of epithelial tissues.

COSMIC (2016) reports *CDH1* mutations in 40 cancer tissue types including stomach (11.40% in N=1342), breast (10.29% in N=3343), large colon (2.87%), skin (2.83%), endometrial (2.81%), and bladder (1.9%) cancer. ICGC reports *CDH1* mutations in 29 cancer tissue types including skin (23.41% in N=598), breast (14.50% in N=1696), ovary (13.98%, N=93), and stomach (11.41% in N=289) cancer samples. *CDH1* mutations are reported at similar rates in breast and stomach cancer in other cancer genomics projects and studies across distinct populations. cBioPortal reports *CDH1* mutation prevalence in stomach cancer at 16.7% (Tokyo Univ., Kakiuchi, 2014, N=30), 15% (Pfizer/UHK, Wang, 2014, N=100), 14.1% (Tianjin Medical University, Chen, 2015, N=78), and 9.4% (TCGA , 2017 prov, N=393). cBioPortal also reports *CDH1* mutation prevalence in breast cancer at 12.7% (TCGA, 2017 prov, N=963) and 10.8% (METABRIC, 2012/2016, N=2051). The rare plasmacytoid bladder cancer subtype also has a high prevalence of *CDH1* mutations in COSMIC (2016) at a rate of 81.8% (N=33). These demonstrate that *CDH1* is important in many cancers and targeting *CDH1* may be widely applied against sporadic cancers in addition to hereditary cancers. However, some of these studies have focused on disease subgroups (such as lobular subtype or estogren receptor negative breast cancers) with poor patient outcomes which may have inflated the prevalence of *CDH1* mutations which are more common in some of these subtypes.

1.3.4.2 Co-occuring Mutations

Another concern is that *CDH1* mutations may co-occur with other known cancer driver genes such as highly prevalent tumour suppressor gene *TP53* or the proto-oncogene *PIK3CA*. cBioPortal reports the prevalence of the mutations in these genes at 10% for *CDH1*, 49% for *TP53*, 22% for *PIK3CA* in stomach cancer (TCGA, 2017 prov, N=393). There is no evidence of significant co-occuring mutations between *CDH1* and *PIK3CA* (mutex $p = 0.231$) but there is evidence for significant mutually exclusive mutations for *CDH1* (mutex $p = 0.002$) and *PIK3CA* (mutex $p = 0.004$) with *TP53*. cBioPortal also reports the prevalence of the mutations in these genes at 13% for *CDH1*, 32% for *TP53*, 36% for *PIK3CA* in breast cancer (TCGA, 2017 prov, N=3963. There is evidence of significant co-occuring mutations with *CDH1* and *PIK3CA* (mutex $p < 0.0001$) and evidence for significant mutually exclusive mutations for *CDH1* (mutex $p = 0.003$) and

PIK3CA (mutex $p = 0.032$) with *TP53*.

These cancer driver mutations have distinct molecular features, leading to disease progression in distinct ways which is a concern for drug resistance when several mutations may accumulate, particularly for sporadic cancers where this is common. Targeting *CDH1* specifically is most suitable for hereditary cancers and combination therapies may be required for sporadic cancers. However, *CDH1* and *TP53* mutant cancers appear to be distinct pathways of tumour progression so the high impact of *TP53* mutation on cancer cells need not be considered for the purposes of studying *CDH1*.

1.3.5 Models of *CDH1* loss in cell lines

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

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

1.4 Summary and Research Direction of Thesis

Genomics technologies and the data made available from them have great potential for understanding of genetics and improving healthcare, including identification of genes altered in cancer for molecular diagnosis, prognostic biomarkers, and therapeutic targets. This has been demonstrated with the identification of cancer genes in many cancers,

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

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

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

To address the concerns raised by recent computational approaches to synthetic lethal discovery in cancer (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015; Wappett *et al.*, 2016), I present similar analysis using solely gene expression data which is widely available for a large number of samples in many different cancers. This uses a statistical methodology the Synthetic Lethal Interaction Prediction Tool (SLIPT) developed for this purpose. To further determine the limitations and implications of synthetic lethal predictions, modelling and simulation was performed upon the statistical behaviour

of synthetic lethal gene pairs in genomics data. Comparison of synthetic lethal gene candidates from public data analysis and experimental candidates, pathway analysis, and networks structure will also be presented to investigate the relationships between synthetic lethal candidates. Release of R codes used for simulation, prediction, and analysis will enable adoption of the methodology in the cancer research community and comparison to existing methods.

My thesis aims to develop such predictions for synthetic lethal partner genes with a focus on the example of E-cadherin to compare to the findings of Telford *et al.* (2015), develop of network approaches for pathway structure, and simulate gene expression on pathway structre with the following bioinformatics and computational biology investigations:

- Developed a query-based synthetic lethal detection methodology (SLIPT) for use on gene expression data
- Adapt this methodology to utilise somatic mutation for query genes or candidate pathway metagenes
- Apply Synthetic lethal prediction to public breast cancer genomics data from TCGA (TCGA, 2012)
- Identify over-represented biological pathways using Reactome (Croft *et al.*, 2014) among synthetic lethal candidate partner genes
- Compare these at the gene and pathway level to experimental screen data in breast cell lines from Telford *et al.* (2015)
- Replicate these analyses in stomach cancer genomics data from TCGA (Bass *et al.*, 2014)
- Determine whether synthetic lethal candidates have importance in biological networks of candidate partner pathways
- Determine whether there are relationships within biological network structures between experimental and predicted gene candidate partners
- Develop a statistical model of synthetic lethal gene expression
- Simulate gene expression with synthetic lethal genes and pathway structures

- Evaluate the effects of modification to the SLIPT procedure on it's statistical performance
- Compare the statistical performance of the SLIPT procedure to alternative statistical methods
- Release a synthetic lethal prediction methodology (SLIPT) to the research community for wider application

Chapter 2

Methods, Techniques, and Resources

In this chapter, I will outline the various existing resources and methods utilised throughout this project. This includes public data repositories, stable and development releases of software packages (mostly for the R programming environment), and implementation of bioinformatics methods 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 to Enable Genomics Research

2.1.1 Public Data and Software Packages

Various bioinformatics resources, such as databases and methods, have become integral parts of genetics and genomics research. Reference genomes, genotyped variants, gene expression, and epigenetics profiles are among the most commonly used resources. Gene expression data in particular is widely available from many microarray and RNA-Seq studies, from repositories such as Gene Expression Omnibus (GEO) (Clough and Barrett, 2016), caArray (Heiskanen *et al.*, 2014), and ArrayExpress (Rustici *et al.*, 2013). Such profiles are excellent resources to examine the changes of gene expression occurring in cancers and the variation between samples. These microarray initiatives have set a precedent for data sharing, data mining, and the wider benefits of publicly

available data for enabling the scientific community to further utilise the data rather than a single research group or consortium (Rung and Brazma, 2013). The practice of integrating findings from publicly available genomics data with the research questions and experimental results of individual research groups has carried over into RNA-Seq datasets including the large-scale cancer genomics projects (Zhang *et al.*, 2011). This thesis is one such example of an investigation enabled by this wider movement and tools developed in various disciplines to generate, process, and disseminate genomic-scale data.

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

2.1.1.1 Cancer Genome Atlas Data

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

Performing a genomic alignment in remains a challenge in bioinformatics and methods to do so may yet be improved (Chen and Tompa, 2010). However, the statistical and biological aspects of bioinformatics are the focus of this thesis, comparing alignment methods is outside the scope of these investigations. The TCGA project (TCGA, 2012) used widely adopted tools: "Bowtie" for alignment (Langmead *et al.*, 2009), "mapslice" to detect splice sites (Wang *et al.*, 2010), and the Reads Per Kilobase per

Million mapped reads (RPKM) approach to qualify reads per transcript as a measure of gene expression (Mortazavi *et al.*, 2008). These are widely acceptable tools for processing RNA-Seq data which were used to produce the raw counts of mapped reads (tier 1) and normalised expression data (tier 3) publicly available from TCGA.

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

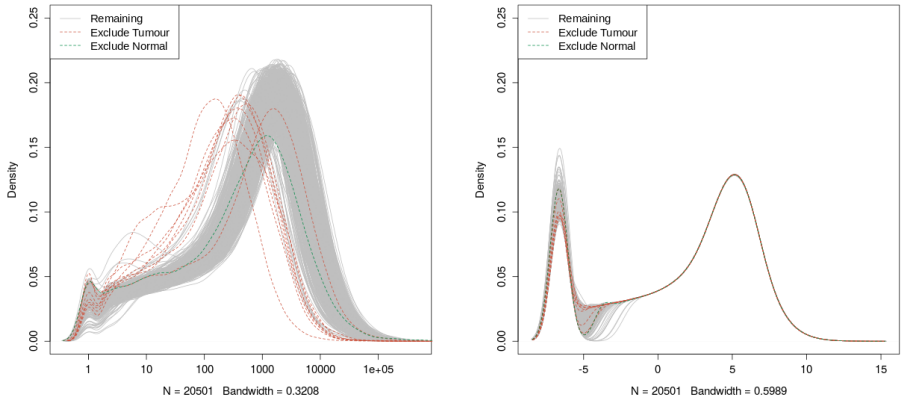
2.1.1.2 Reactome and Annotation Data

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

2.2 Data Handling

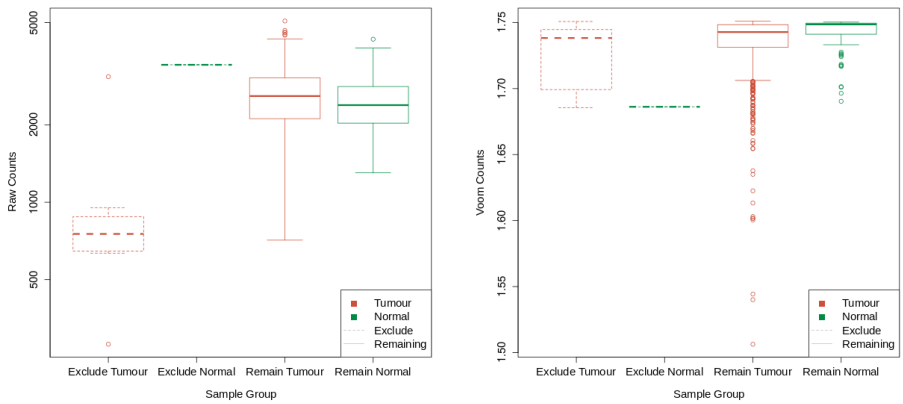
2.2.1 Normalisation (voom)

Apart from the PAM50 subtyping procedure (Parker *et al.*, 2009), which required RSEM normalised data (J.S. Parker personal communication), the analysis of the RNA-Seq data presented here was based on raw read count data. Raw read counts were log-scaled; samples removed for consistency (based on a Euclidean distance correlation matrix as described in section 2.2.2); and the final dataset was TMM normalised (Robinson and Oshlack, 2010) then processed using the `voom` function (Law *et al.*, 2014) in the `limma` R package (Ritchie *et al.*, 2015).



(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.

2.2.2 Sample Triage

The TCGA RNA-Seq data were assessed for batch effects using a correlation matrix of the log-transformed raw counts for which a heatmap (Euclidean distance, complete linkage) is shown in Figure A.2. While no major batch effects were detectable between

the samples, 9 samples were excluded due to poor correlation with the remaining samples, as detailed in Table 2.1. These samples showed unusual density plots compared to the rest of the dataset, and exhibited low mean read count in Figures 2.1 and 2.2. A heatmap showing key clinical properties of these excluded samples and their correlation with the remainder of the samples is shown in Figure A.1, and a full correlation heatmap (Figure A.2) shows these samples as relatively poorly correlated outliers in the bottom rows and left columns. After removal of these samples, the TCGA dataset used for analysis consisted of the remaining 1168 samples (from 1040 patients): 1049 tumour samples, 112 normal tissue for matched samples, and 7 metastases.

Table 2.1: Excluded Samples by Batch and Clinical Characteristics.

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

2.2.3 Pathway Metagenes and the Singular Value Decomposition

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

2.2.3.1 Candidate Triage and Integration with Screen Data

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

The pathway composition of the intersection was further verified by a permutation resampling analysis (as described in section 2.3.6): the same number of genes detected by SLIPT were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1,652 reactome pathways. These over-representation scores (χ^2) were compared the observed over-representation in the intersection of the SLIPT candidates, with the proportion of resamplings with higher χ^2 values used for empirical p-values of pathway composition. Pathways for which no resamplings were occurred as high as the observed were reported as $p < 10^{-6}$. These empirical p-values were adjusted for multiple comparisons (FDR). Intersection size was not assumed to be constant across resamplings so similarly with the proportion of resamplings with higher or lower intersection size were used to evaluate significance of enrichment or depletion respectively (of siRNA candidate among SLIPT candidate genes).

2.3 Techniques

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

2.3.1 Statistical Procedures and Tests

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

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

This generates data where each row (gene) has a mean of 0 and standard deviation of 1. Where plotted as 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.*, 2015). Where specified, the Fisher's exact test, χ^2 test, and correlation were used to measure associations between variables (as implemented in the `stats` R package (R Core Team, 2016)). Unless otherwise specified, Pearson's correlation was used for correlation analyses (r) and coefficient of determination (R^2). Where these comparisons are discussed in more detail, Fisher's exact test and χ^2 tests are supported by a table or Venn diagram, rendered with the `limma` R package (Ritchie *et al.*, 2015). In some analyses, correlation is 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, 2016) has also been used to implement the t-test to compare pairs of data. Where relevant, an analysis of variance (ANOVA) has been performed to report significance of multivariate predictors of outcomes, or least squares linear regression performed for the adjusted coefficient of determination (R^2) and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violinplot visualisation (rendered in R).

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

2.3.2 Gene Set Over-representation Analysis

Gene set enrichment over-representation 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) or comparison with experimental candidates (see section 2.2.3.1). Initially, these tests were performed using the GeneSetDB web tool (Araki *et al.*, 2012) hosted by the University of Auckland on the Reactome pathways (Croft *et al.*, 2014). Since the GeneSetDB tool used an older version of Reactome (version 40), it was difficult to directly compare with the results of other analysis (see sections 2.2.3.1 and 2.3.6) performed on version 52 (as described in section 2.1.1.2). Thus an implementation of the hypergeometric test in R (R Core Team, 2016) was used to test for over-representation against Reactome (version 52) pathways. Pathways containing less than 10 genes or more than 500 (as performed in GeneSetDB by Araki *et al.*, 2012) were excluded before adjusting for multiple comparisons.

2.3.3 Clustering

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

2.3.4 Heatmap

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

2.3.5 Modeling and Simulations

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

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

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for χ^2, significance and directional criteria met for SLIPT (see section 3.1), and an arbitrary threshold: < -0.2 and $> +0.2$ for negative correlation and 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, 2007; Fawcett, 2006; Zweig and Campbell, 1993). These curves (implemented with the `ROCR` R package (Sing *et al.*, 2005)) plot the True Positive Rate (sensitivity) against the False Positive Rate (1−specificity) as the prediction threshold is varied. An ideal detection method will have a true positive rate of 1 and a false positive rate of 0, hence the Area Under the ROC curve (AUC or AUROC) is a measure of statistical performance for a detection procedure accounting for this trade-off. AUROC values are typically range from 0.5 the value expected by random chance to 1 for an optimal detection method, however it is possible for an AUROC below 0.5 for a poor detection method that performs worse than random chance. An cancer biology, an AUROC of approximately 0.8 is a predictive biomarker suitable for publication (Hajian-Tilaki, 2013) but predictors with lower AUROC values may still be informative depending on the context. In this thesis, the AUROC values varies widely across simulation parameters and a primarily used for comparisons across these parameters, although they can also be used to refine thresholds for optimal classification.

2.3.6 Resampling Analysis

Resampling analyses (e.g., "permutation" analysis) are used to statistically test the significance of an observation without assuming the underlying distribution of expected test statistics Collingridge (2013). Instead these are derived from randomly shuffling test statistics or randomly sampling predicted candidates. For the purposes of this thesis, this involved randomly sampling genes from those tested to be analysed as putative synthetic lethal candidates. This was performed both for testing the significance of pathway composition in the intersection with experimental gene candidates (section 2.2.3.1) and for assessing the significance of pathway structure among synthetic lethal candidates (section 3.4.1.1).

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

these iterations that are greater than or equal to that observed, forms an empirically derived p-value to test significance.

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

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

2.4 Pathway Structure Methods

2.4.1 Network and Graph Analysis

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

Graph theory is a branch of pure mathematics which deals with the properties of

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

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

2.4.2 Sourcing Graph Structure Data

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

2.4.3 Constructing Pathway Subgraphs

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

2.4.4 Network Analysis Metrics

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

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

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

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

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

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

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

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

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

2.5 Implementation

2.5.1 Computational Resources and Linux Utilities

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

Table 2.2: Computers used during Thesis

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

Table 2.3: Linux Utilities and Applications used during Thesis

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

Table 2.4: R Installations used during Thesis

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

2.5.2 R Language and Packages

The R programming language has been used for the majority of this thesis. Current R installations across the machines used are given in Table 2.4. Local machines currently run the latest version of the R (at the time of writing) and remote machines run the versions and modules as managed by the system administrator. Various scripts and packages in this thesis were developed or run in previous versions of RStudio and R but these run without error in the current version of R (and the older versions on remote machines). The R packages developed during this thesis are given in Table 2.5 with the relevant sections describing their implementation and use where appropriate. Various

R packages were used throughout this thesis (as detailed in Table 2.6 with versions specified), which 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 the Appendix (Table B.1).

Table 2.5: R Packages Developed during Thesis

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

Table 2.6: 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

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, 2017). The Pan cluster used throughout this thesis project to optimise and perform computations which would have otherwise been infeasible in the timeframe of thesis. Such technological developments and infrastructure initiatives have enabled bioinformatics research including this project. High performance computing on the Pan cluster was used extensively in this project including for resampling analysis (in sections 2.3.6 and 3.4.1.1), calculating information centrality (in section 2.4.4), and in simulations (in sections 2.3.5, 3.2, and 3.4.2)

Scripts and data were transferred between the Pan cluster and University of Otago computing resources by `rsync` or the Globus file transfer service (Globus, 2017). R scripts (R Core Team, 2016) were run in parallel with the "simple network of workstations" `snow` R package Tierney *et al.* (2015). This utilised the "message passing interface" (Yu, 2002) when it was feasible with memory requirements to run in parallel across multiple compute nodes, otherwise 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, 2017). When running R scripts across many parameters or for memory-intensive jobs, Slurm array job submission and independent submission of different parameters via shell commands with arguments passed to R. In some cases, this submission was automated across a range of parameters with Bash scripts.

Chapter 3

Methods Developed During Thesis

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

3.1 A Synthetic Lethal Detection Methodology

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

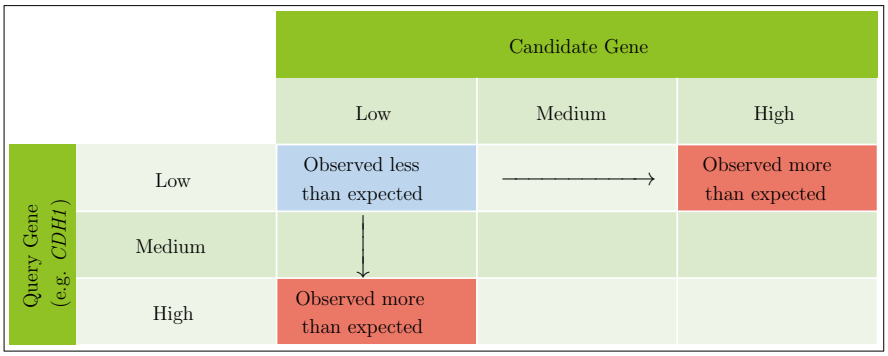


Figure 3.1: **Framework for synthetic lethal prediction.** Synthetic Lethal Interaction Prediction Tool (SLIPT) was designed to identify candidate interacting genes from gene expression data using the χ^2 test against a query gene. Samples are sorted into low, medium, and high expression quantiles for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

are less frequent than expected; and (iii) observed low-query, high-candidate and high-query, low-candidate samples are more frequent than expected.

The synthetic lethal prediction procedure has also been adapted to utilise somatic mutation data for the query gene. This is intended to utilise a query gene known to be recurrently mutated in the disease (and dataset), with the majority of mutations inactivating gene function (such as null or frameshift mutations). A synthetic lethal interaction is predicted (as shown in Figure 3.2) when (i) the χ^2 test is significant; (ii) observed mutant-query, low-candidate samples are less frequent than expected; and (iii) observed mutant-query, high-candidate and wild-type-query, low-candidate samples are more frequent than expected. Unless otherwise specified, computationally predicted synthetic lethal gene candidates from SLIPT used expression data (exprSL) for both genes (as shown in Figure 3.1) rather than mutation data (mtSL) for the query gene (as shown in Figure 3.2).

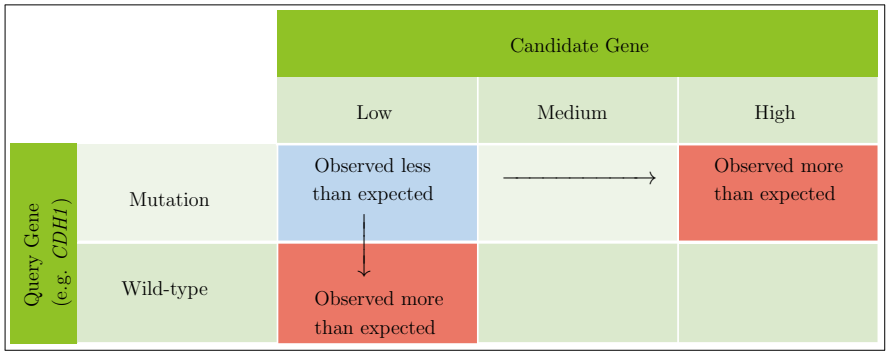


Figure 3.2: **Synthetic lethal prediction adapted for mutation.** Synthetic Lethal Interaction Prediction Tool (SLIPT) was also adapted to identify candidate interacting genes using (somatic) mutation data of the query gene in the χ^2 test. Samples are sorted into low, medium, and high expression quantiles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

3.2 Synthetic Lethal Simulation and Modelling

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

3.2.1 A Model of Synthetic Lethality in Expression Data

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

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

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

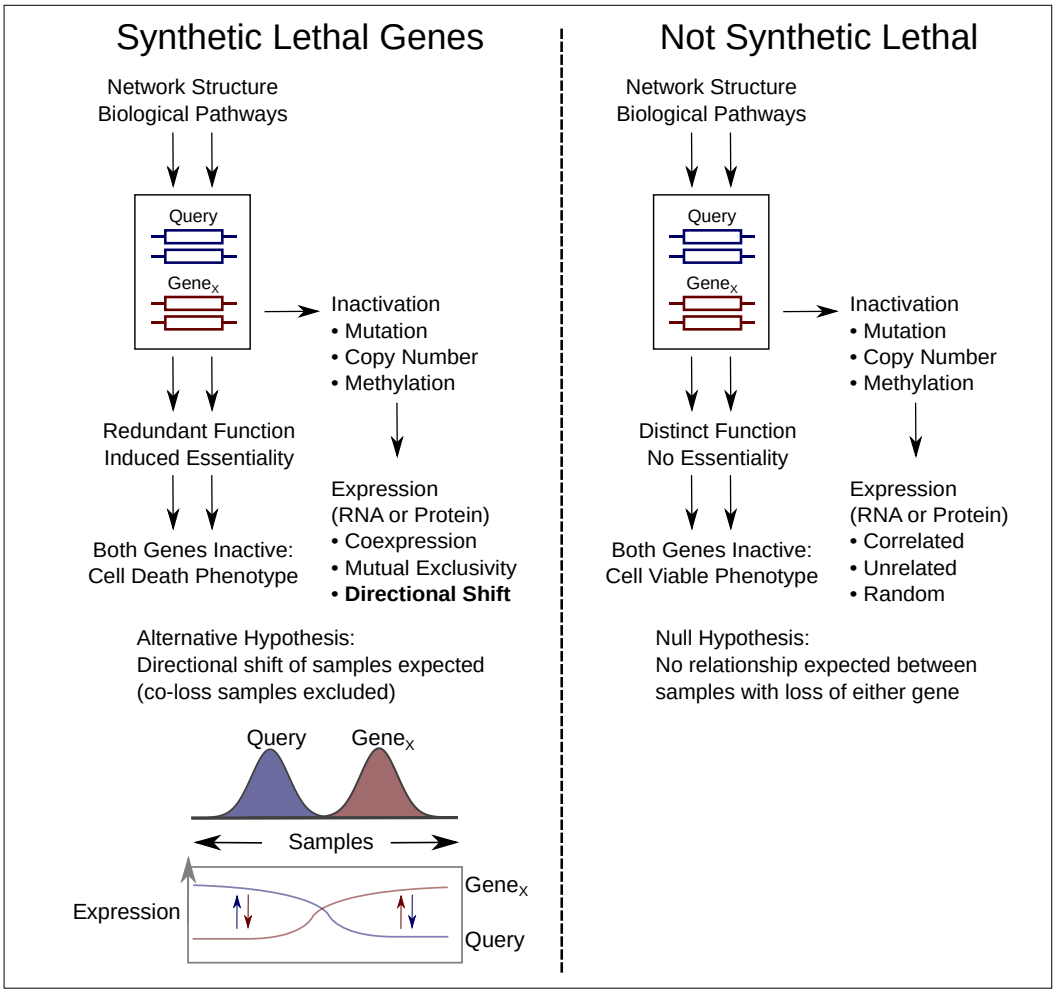


Figure 3.3: **A model of synthetic lethal gene expresion.** A conceptual model of synthetic lethal interactions between a Query gene and partner gene (G_X). Genes that are synthetic lethal may not both be non-functional in the same sample without another gene compensating for the loss of function. This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

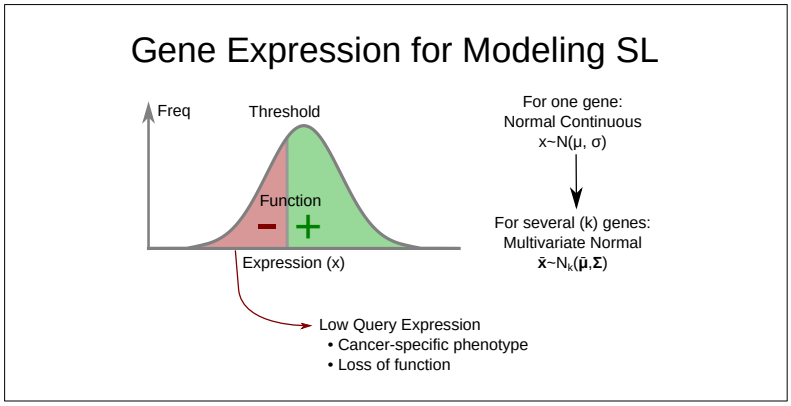


Figure 3.4: **Modeling synthetic lethal gene expression.** When modeling synthetic lethal interactions between a Query gene with partner genes (G_X and G_Y) above, cellular viability requires that at least of genes is not inactivated. Expression below a threshold is used as a model of loss of function,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

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

A synthetic lethal pair of genes is unlikely to act in isolation, therefore higher-order synthetic lethal interactions (i.e., 3 or more genes) must be considered in the model as shown in Figure 3.5. Even when testing pairwise interactions, modelling higher level interactions that may interfere is important. If there are additional synthetic lethal partners, there are two possibilities for adding these: 1) that they are independent

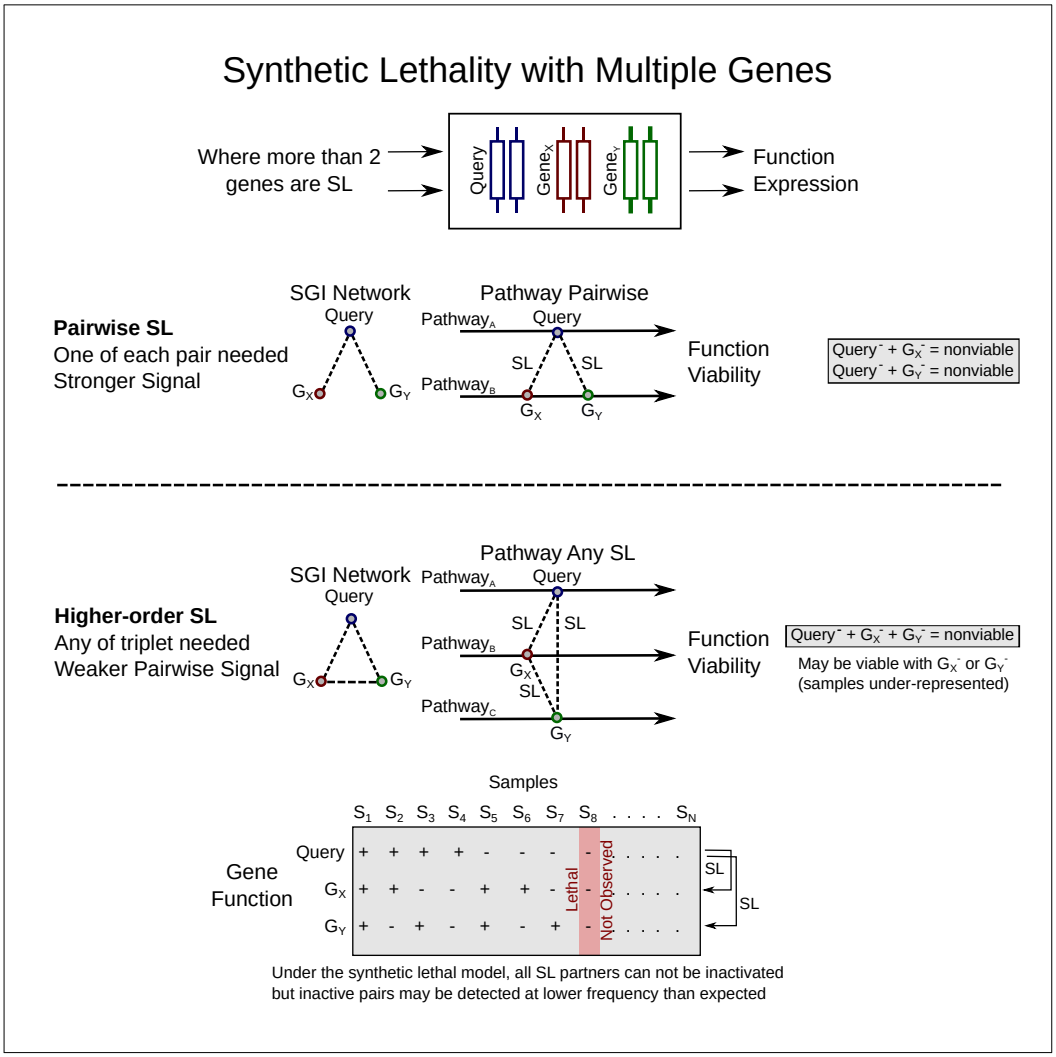


Figure 3.5: **Synthetic lethality with multiple genes.** Higher order synthetic lethal interactions may occur between 3 or more genes, affecting the simulated expression (or synthetic lethal predictions) even if undetected when observed pairwise. Consider interactions between a Query gene and two partner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

partners of the query genes interacting pairwise (and not with each other) or 2) that an addition partner gene interacts with both of the synthetic lethal genes already in the system and any of the three (or more) are required to be functional for the cell to survive.

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

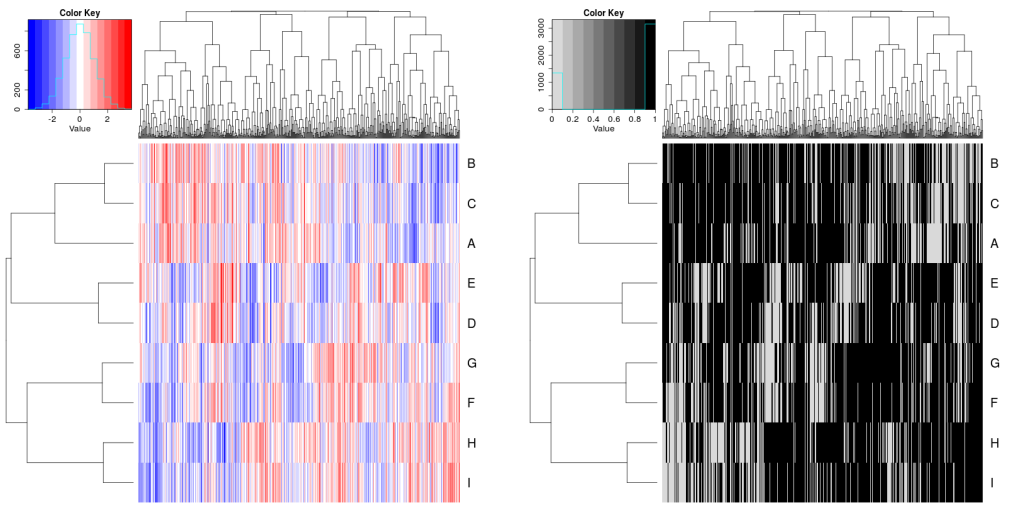
In the example of 3 synthetic lethal genes 3.5, only one of genes involved in the higher-order synthetic lethal interaction is required for cell viability. For synthetic lethal pairs, only a subset of these samples will be inviable (i.e., removed from simulated data), leading to an under-representation.

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

3.2.2 Simulation Procedure

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

Sampling a distribution for expression profiles has the added advantage of being amenable to simulating correlation structures with the multivariate normal distribu-



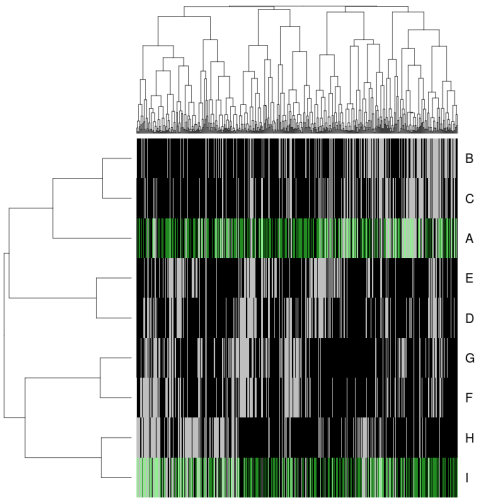
(a) Simulated expression matrix (b) Corresponding gene function calls

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

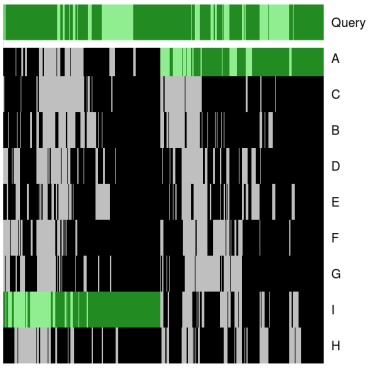
tion (using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016)). The parameter Σ is a covariance matrix defines the correlation structure between simulated genes being sampled. With a diagonal of one, this Σ matrix simulates genes with a standard deviation of one and the covariance parameters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

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

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

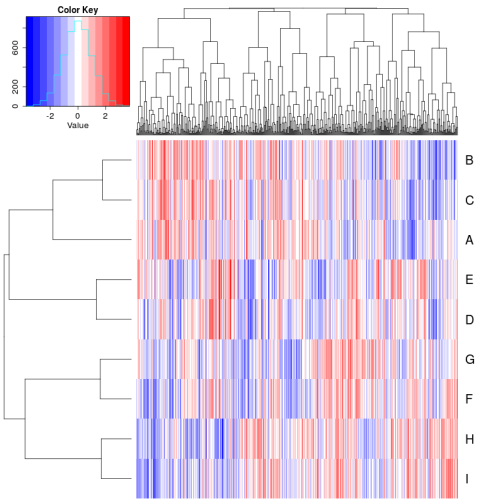


(a) Simulated gene function with SL genes

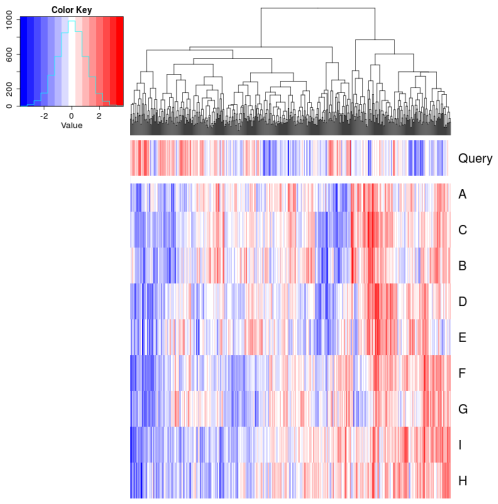


(b) Query gene added with SL condition

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



(a) Initial expression matrix



(b) Simulated synthetic lethal dataset

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

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

3.3 Detecting Simulated Synthetic Lethal Partners

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

3.3.1 Binomial Simulation of Synthetic lethality

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

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

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

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

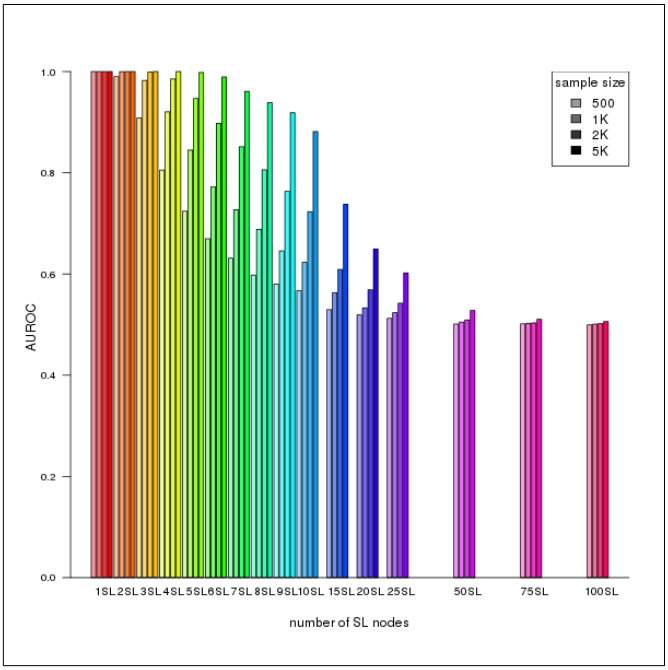


Figure 3.9: **Performance of binomial simulations.** Gene function was simulated by binomial sampling and tested for synthetic lethal genes. Statistical performance declines with additional known synthetic partners but this is mitigated by increased sample sizes.

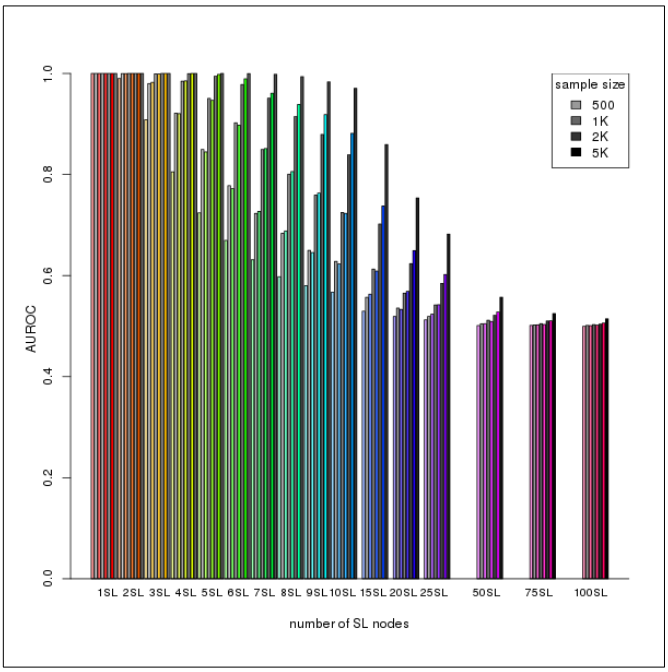


Figure 3.10: **Comparison of statistical performance.** Binomial simulation of synthetic lethality (in colour) is compared (in greyscale) to multivariate normal simulations (detailed below) which consistently outperforms binomial simulation across parameters.

extent, as expected with a statistical predictor, particularly for moderate numbers of synthetic lethal partners.

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

The multivariate normal simulation procedure is more representative of the (normalised) expression data SLIPT is intended for and enables the prediction procedure to be tested without changes to the methodology (presented in more detail in section 3.3.2). Sampling continuous expression values from a normal distribution allows the expression threshold for gene function to differ from the categorical "low" and "high" expression binning performed by SLIPT (as discussed in section 3.2.1) which represents that the SLIPT procedure does not assume a known threshold for expression but rather uses expression as an estimate of gene function. This functionality can be included in the multivariate normal simulation without compromising the statistical performance of the SLIPT, rather the performance estimates (shown in Figure 3.10) were a marked improvement over the binomial simulation procedure across simulation parameters in an equivalent simulation (without correlation structure).

3.3.2 Multivariate Normal Simulation of Synthetic lethality

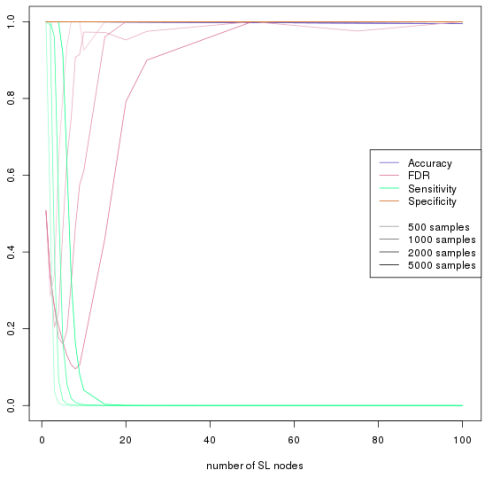
The multivariate normal simulation procedure was initially performed using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016) (as described in section 3.2) without correlation structure.

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

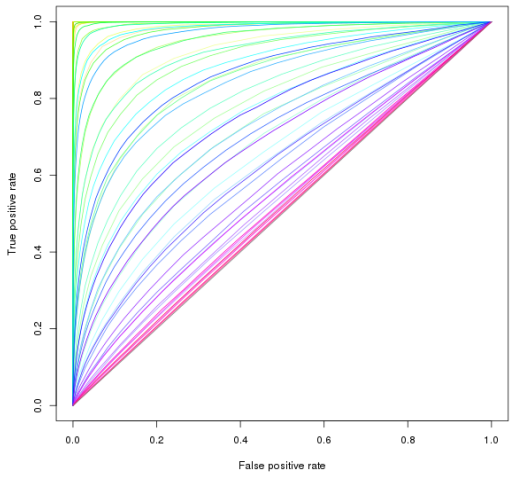
$$X \sim N(\mu, \Sigma)$$

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

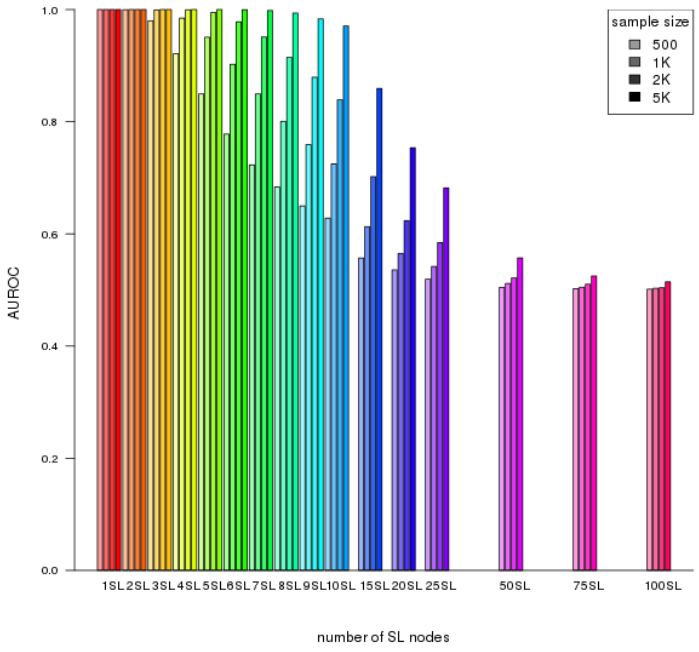
As shown in Figure 3.11(a), the statistical accuracy of SLIPT as a binary classifier is considerably high across simulations of a full human dataset of 20,000 genes. However,



(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Statistical performance

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

with the χ^2 p-value as a threshold for prediction, this is largely to desirable specificity: the majority of non-SL genes are distinguished from the few underlying synthetic lethal genes. In this regard, the SLIPT methodology generally performs better with larger datasets with more expected negatives and thus the results of simulations of smaller numbers of genes (such as the graph structures analysed in Chapter 6) can be applied to larger datasets where they are expected to perform comparably or better with a lower false negative rate. Accordingly, key results will be supported by replication with larger numbers of non-SL genes added to the simulations.

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

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

The receiver operating characteristic curves (in Figure 3.11(b)) demonstrate that SLIPT is subject to near equal trade-off between sensitivity and specificity across threshold values. The lower sensitivity and higher specificity with a binary classifiation (in Figure 3.11(a)) stems from stringent testing by SLIPT with (FDR) p-values adjusted for multiple tests. The area under these curves is also used to compare statistical

performce (in Figure 3.11(c)), with declining performance across increased underlying synthetic lethal partners and increased performance with sample size in multivariate normal simulations.

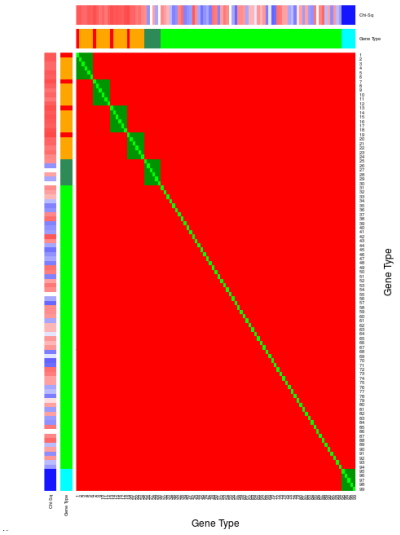
3.3.2.1 Multivariate Normal Simulation with Correlated Genes

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

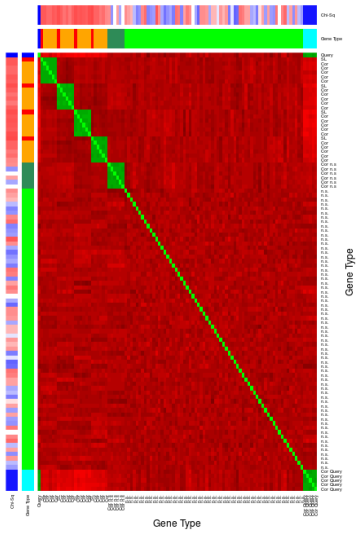
The use of correlation structures generalises to larger datasets, such as 1000 genes shown in Figure 3.13. Synthetic lethal genes are highly ranked by SLIPT and still largely distinguishable from correlated genes. As previously discussed in section 3.3.2, these synthetic lethal genes are still detectable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

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

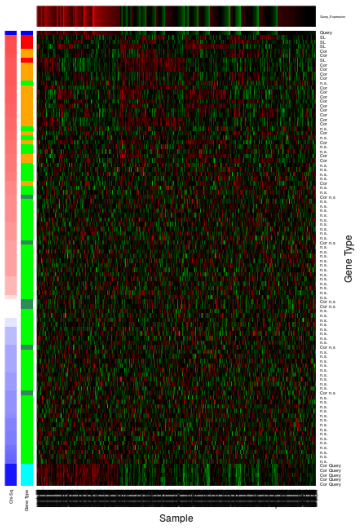
These simulations (on 100 genes) were repeated to examine the variation between detection on different samples and varying the number of underlying synthetic lethal partners, in simulated gene expression data with correlations structure. A small nuber (10 for each) simulations are shown in Figure 3.14 to demonstrate the variation be-



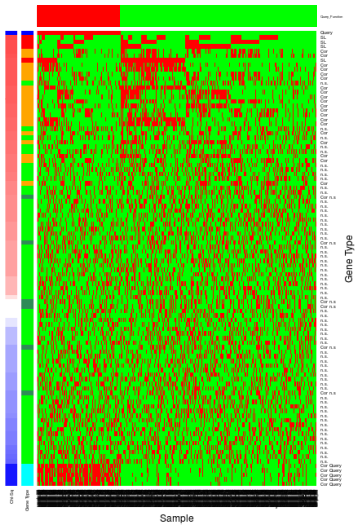
(a) Input Σ matrix parameter



(b) Simulated correlation matrix

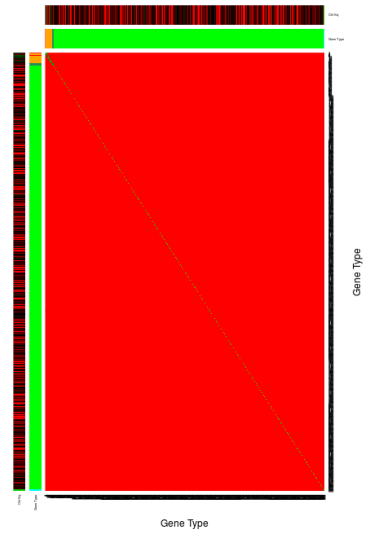


(c) Simulated gene expression

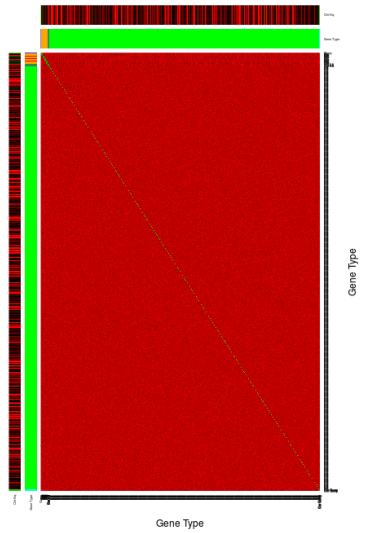


(d) Simulated gene function

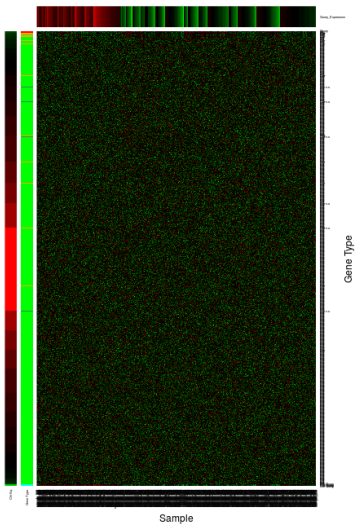
Figure 3.12: **Simulating expression with correlated gene blocks.** A Σ matrix (a) is used generate a multivariate normal distribution with of 100 genes correlated blocks of genes (correlated by 0.8) with a comparable structure (b) to the input Σ, as shown by correlation on a red–green scale. The annotation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



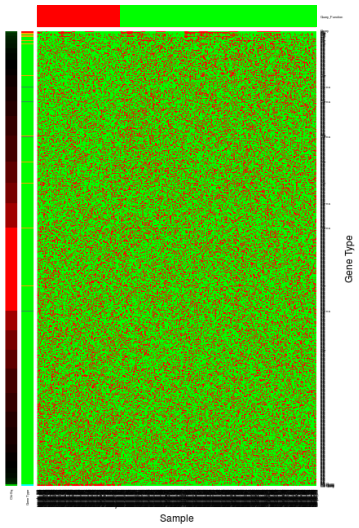
(a) Input Σ matrix parameter



(b) Simulated correlation matrix

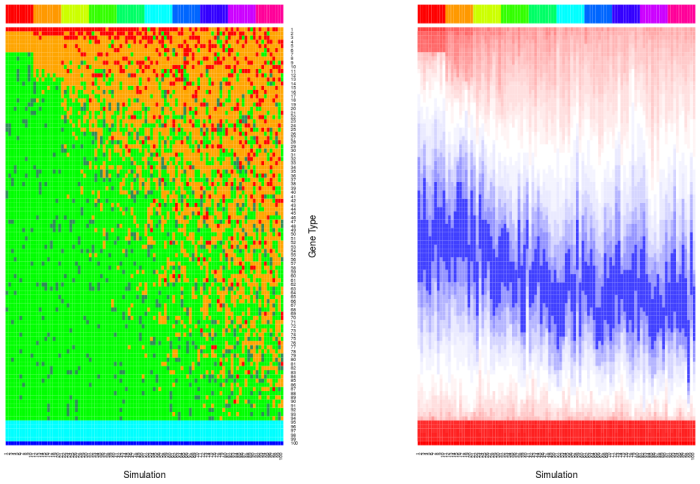


(c) Simulated gene expression



(d) Simulated gene function

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

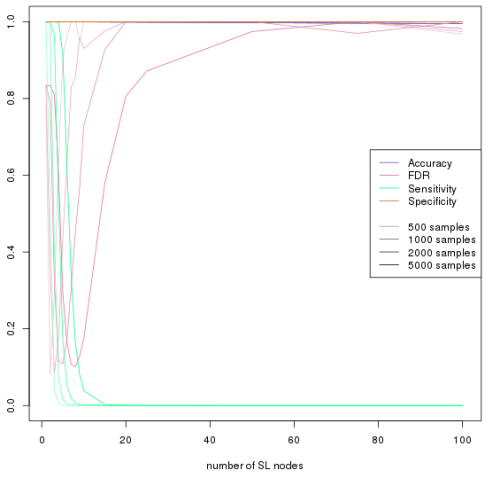


(a) Gene category in simulations (b) Corresponding χ^2 values

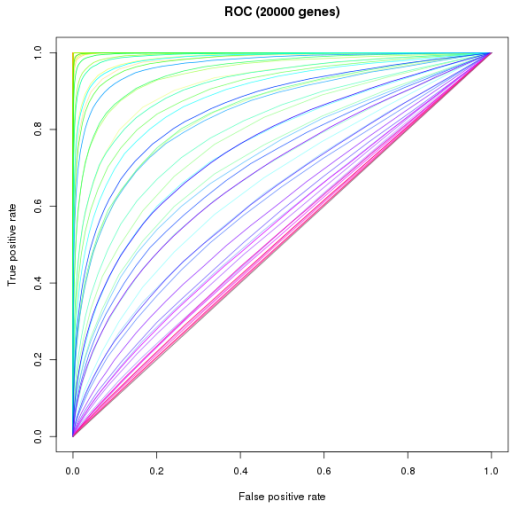
Figure 3.14: **Synthetic lethal prediction across simulations.** The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ^2 signed by the SLIPT directional condition is shown across simulations.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

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

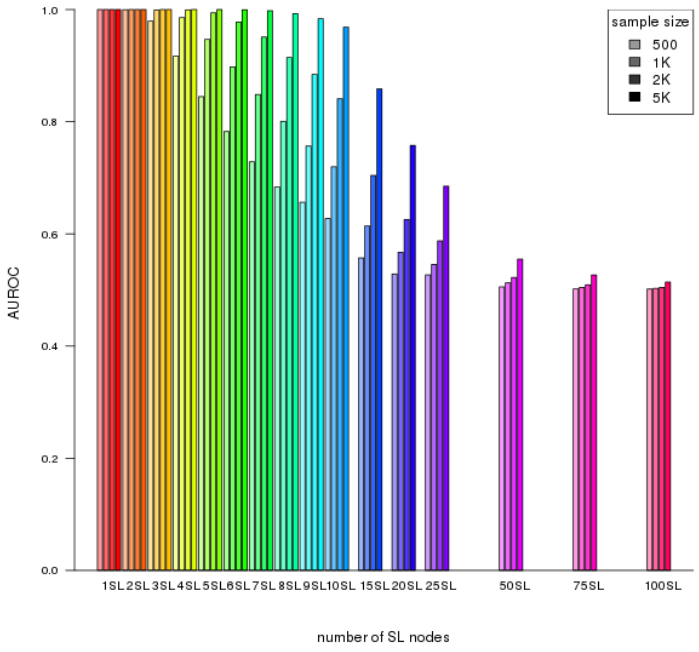
Whether the synthetic lethal genes detected in simple simulations (in Figure 3.14) are robustly detectable across greater number of simulations, in addition to further comparions, was tested with a supporting ROC analysis. These results (in Figure 3.15) are very similar to simulations without correlation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



(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Statistical performance

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

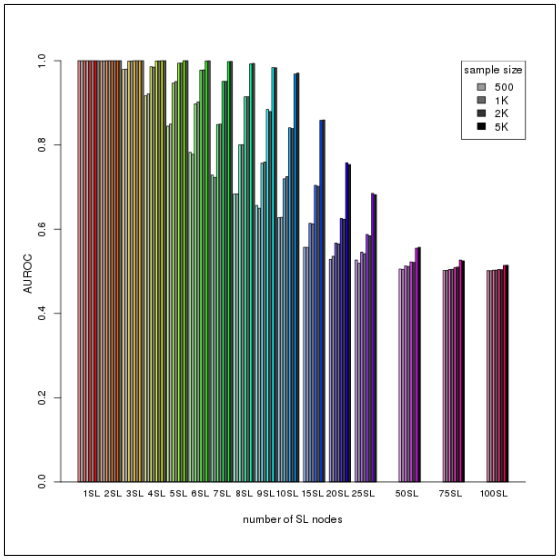
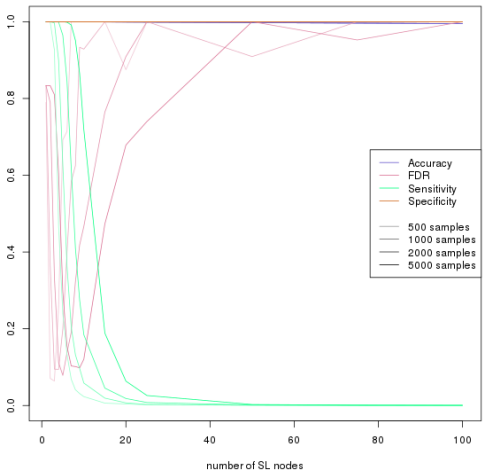


Figure 3.16: **Comparison of statistical performance with correlation structure.** Multivariate simulation of synthetic lethality with correlation structure (in colour) has comparable performance to simulation without correlations (in greyscale) with known synthetic partners across parameters.

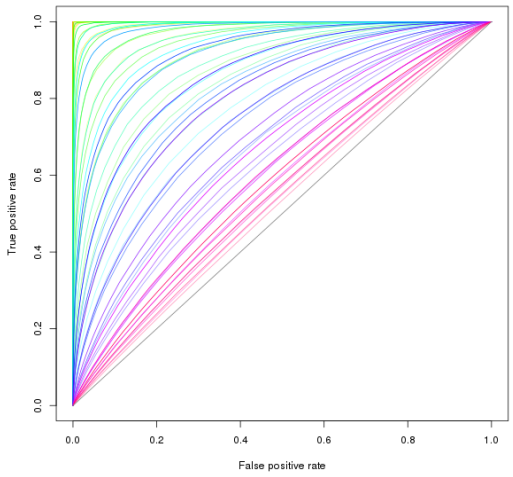
negatives. This is reflected in a similar decline in statistical performance for increasing numbers of synthetic lethal partners and a compensating increase in performance with higher sample size. Overall, the statistical performance is very similar to simulations without correlation structure (as shown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

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

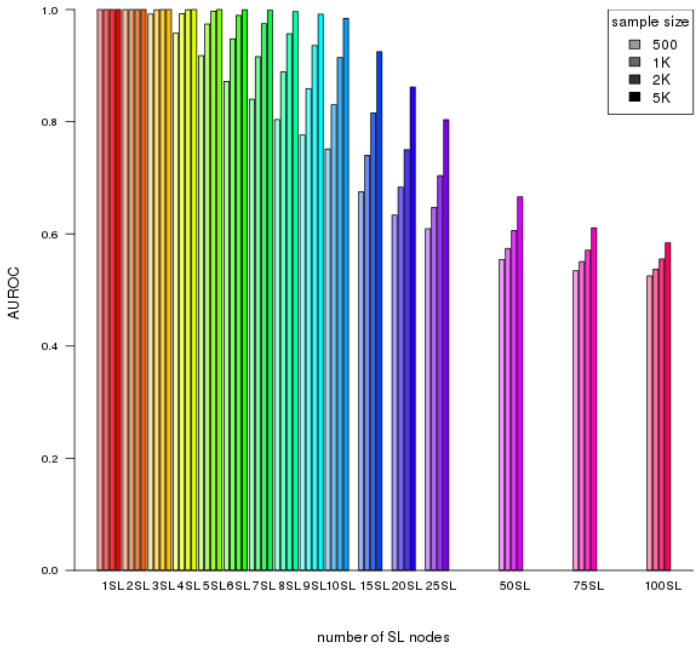
In particular, genes correlated with true synthetic lethal genes have little impact on the performance of SLIPT detection: synthetic lethal genes are as distinguishable from true negative genes as without correlated genes. Synthetic lethal correlated genes will not interfere detect of true synthetic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



(a) Statistical evaluation



(b) Receiver operating characteristic



(c) Statistical performance

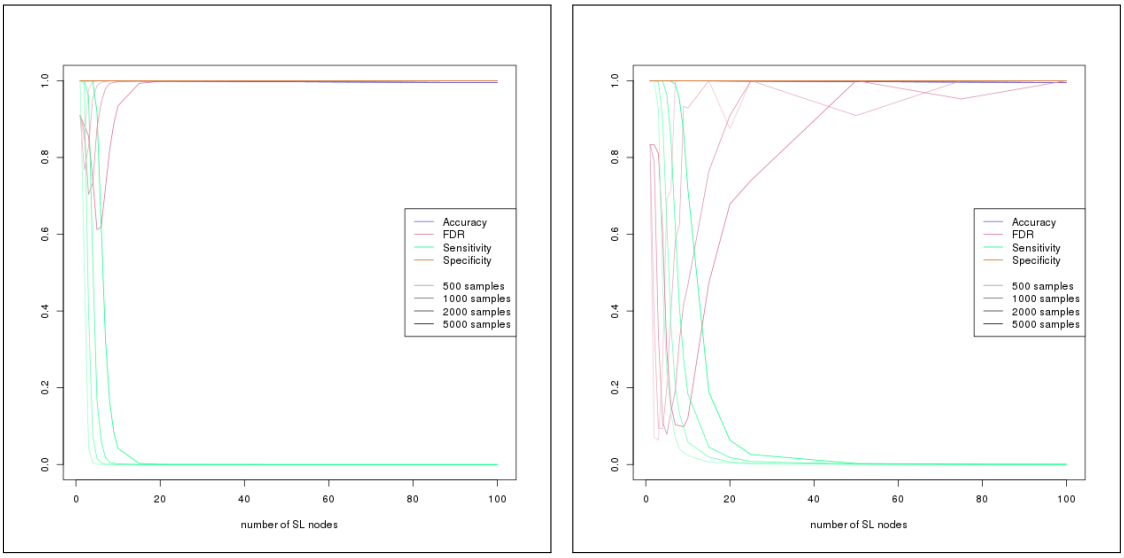
Figure 3.17: **Performance with query correlations.** Simulation of synthetic lethality was performed sampling from a multivariate normal distribution (with correlation structure including correlated genes with non-SL and query genes). As before, performance of SLIPT declines for more synthetic partners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

3.3.2.2 Specificity with Query-Correlated Pathways

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

3.3.2.2.1 Importance of Directional Testing

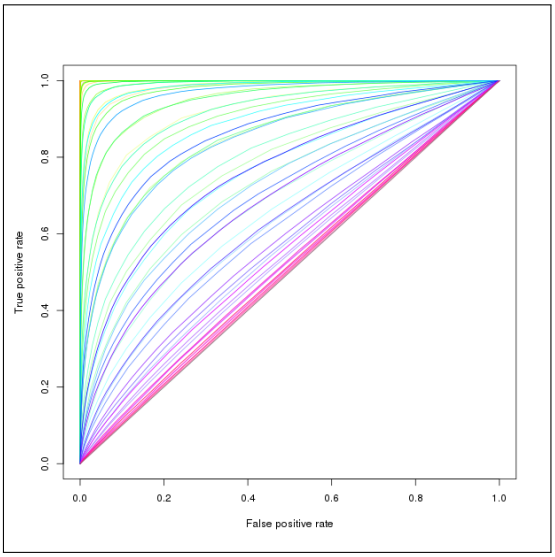
It is important to notice here that the directional criteria of the SLIPT procedure is enhancing it's performance, particularly in distinguishing positively correlated true negatives. The multivariate normal simulation results, with 20,000 genes including all of the correlation structures discussed (SL,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



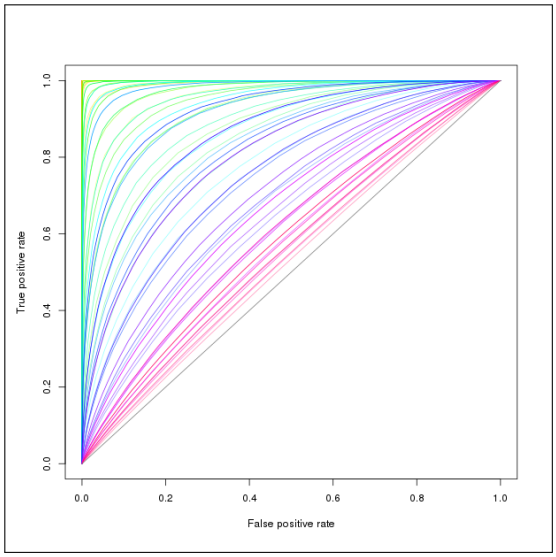
(a) χ^2 testing without direction

(b) SLIPT with directional criteria

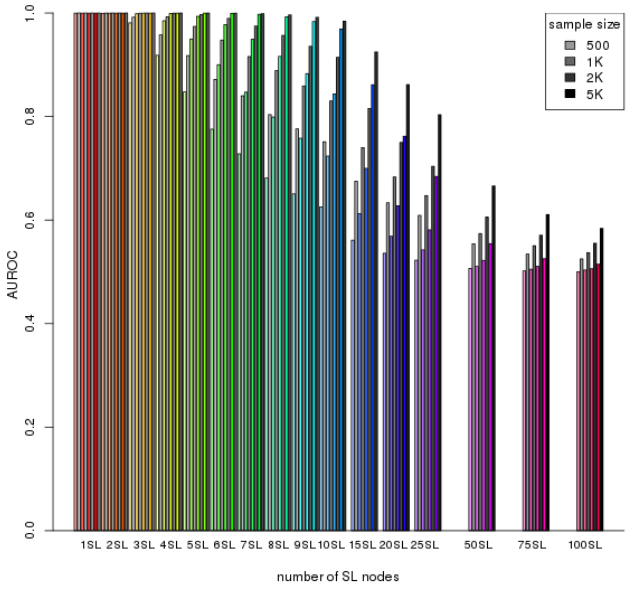
Figure 3.18: **Statistical evaluation of directional criteria.** A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and and the equivalent χ^2 test without. SLIPT exhibited a consistently higher sensitivity and lower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



(a) χ^2 testing without direction



(b) SLIPT with directional criteria



(c) Statistical performance

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

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

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

3.4 Graph Structure Methods

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

3.4.1 Upstream and Downstream Gene Detection

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

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

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

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

3.4.1.1 Permutation Analysis for Statistical Significance

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

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

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

counts procedure that excluded them. A refined version of this procedure ensured that the number of intersecting genes was equal to the number observed to test for pathway structure without changing the intersection size, the subject of prior analyses.

3.4.1.2 Ranking Based on Biological Context

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

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

$$rank(z) = 1 + diameter(G)$$

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

$$rank(i) = rank(z) - d_{iz}$$

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

$$rank(j) = rank(i) + d_{ij}$$

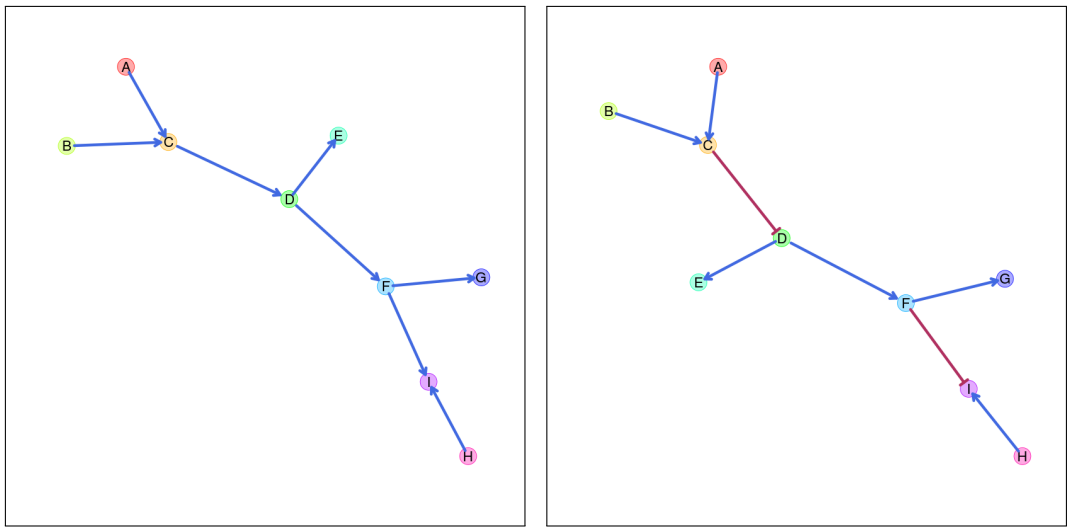
This process may be performed iteratively to fill in pathway ranking but it was not necessary to perform further iterations for the candidate synthetic lethal pathways investigated (amenable to this procedure) which exhibited clear directional structure and the small world property (with a low diameter).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

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

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

3.4.2 Simulating Gene Expression from Graph Structures

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



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

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

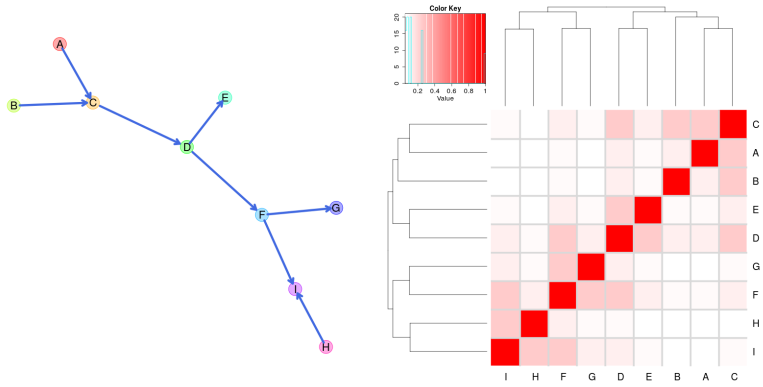
First, the graph structure is constructed for simulated data to be generated from (by sampling from a multivariate normal distribution using the `mvtnorm` R package (Genz and Bretz, 2009; Genz *et al.*, 2016)). Throughout this section, the simulation procedure will be demonstrated with the relatively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

and `igraph.extensions` library (see Table 2.5 and section 3.5.3). The `plot_directed` function allows customisation of plot parameters for each node or edge and mixed (directed) edge types for indicating activation or inhibition. These inhibition links (which often occur in biological pathways) are demonstrated in Figure 3.20(b).

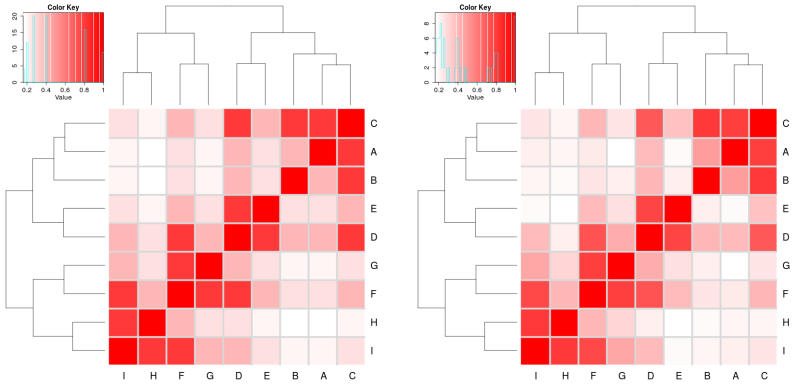
The simulation procedure is designed to use such graph structures to inform development of a "Sigma" variance-covariance matrix (Σ) for sampling from a multivariate normal distribution (using the `mvtnorm` R package). Given a graph structure (or adjacency matrix), such as Figure 3.21(a), a relation matrix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

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

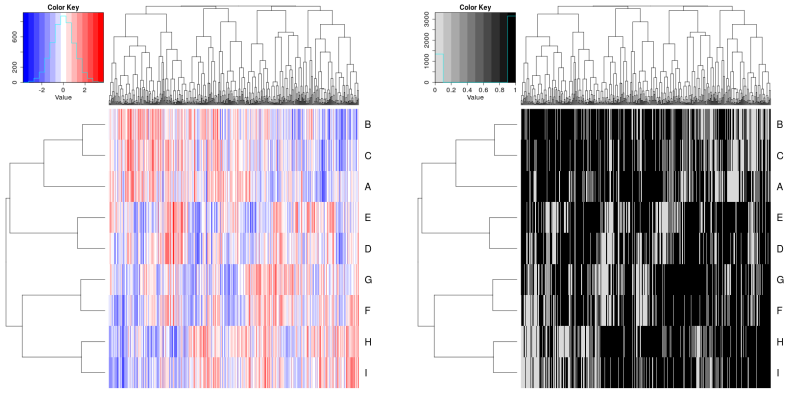
The simulation procedure (depicted in Figure 3.21) is amenable to pathways containing inhibition links (as shown in Figure 3.22) with several refinements. With the inhibition links (as shown in Figure 3.22(a)), distances are calculated in the same man-



(a) Activating pathway structure (b) Distance matrix

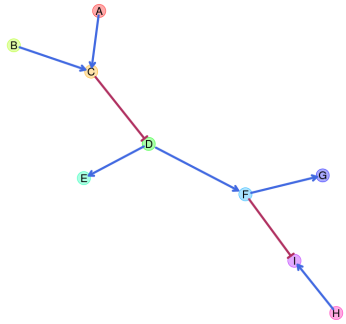


(c) Sigma, Σ (expected correlation) (d) Simulated correlation structure

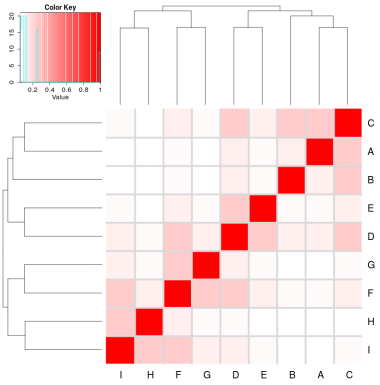


(e) Simulated expression data (f) Simulated gene function calls

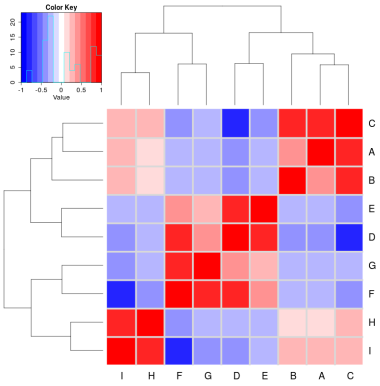
Figure 3.21: **Simulating expression from a graph structure.** An example graph structure is used to derive a correlation structure from the relative distances between nodes and simulate continuous gene expression with sampling from the multivariate normal distribution.



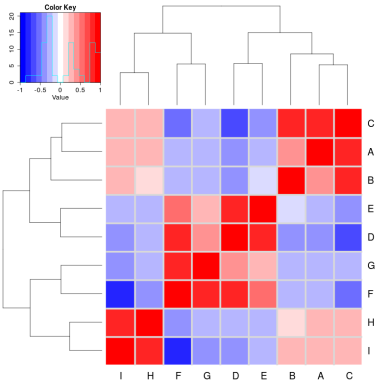
(a) Pathway structure with inhibition



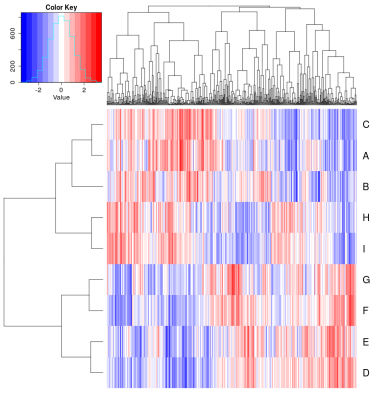
(b) Distance matrix



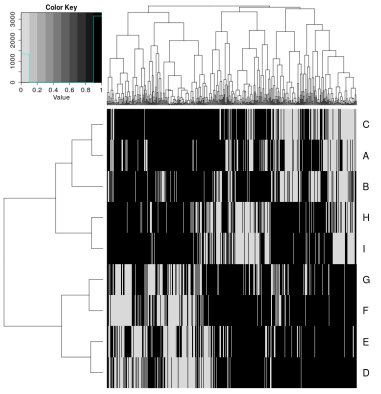
(c) Sigma, Σ (expected correlation)



(d) Simulated correlation structure



(e) Simulated expression data



(f) Simulated gene function calls

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

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

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

3.5 Customised Functions and Packages Developed

[Move to Appendix?]

Various R packages have been developed throughout this thesis using `devtools` (Wickham and Chang, 2016) and `roxygen` (Wickham *et al.*, 2017) to enable reproducibility of customised analysis and visualisation. Many of these have the added benefit of the functions being documented, demonstrated in example vignettes, and released on GitHub to enable the research community to access utilise them in their own analysis. These are summarised in Table 2.5, along with the corresponding urls for their GitHub repository which contains a README file with instructions for installation with the `devtools` R package (Wickham and Chang, 2016) and links to the relevant vignette(s) where available.

3.5.1 Synthetic Lethal Interaction Prediction Tool

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

SLIPT is amenable to analysis of any effectively continuous measure of gene activity (e.g., microarray, RNA-Seq, protein abundance, or pathway metagenes). Executing `slipt` is straightforward: the `prep_data_for_SL` function scores samples as "low", "medium", or "high" for each gene, then the `detect_SL` function tests a given query gene against all potential partners by performing the chi-squared test and directional

conditions. This function returns a table summarising the observed and expected sample numbers used for the directional criteria, the χ^2 values, and corresponding p-values including adjusting for multiple comparisons. The `count_of_SL` and `table_of_SL` functions serve to facilitate summary and extraction of the positive SLIPT hits, respectively, from the table of predictions of synthetic lethal partners.

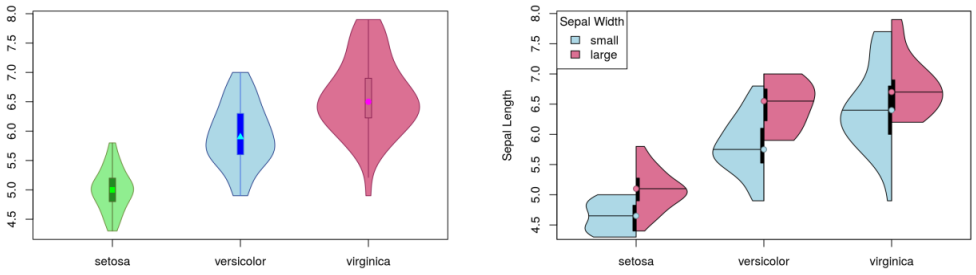
The SLIPT methodology in this package release has been used in later analyses rather than the corresponding source R code, including use on remote machines and upon simulated data. In particular, the functions in the package facilitate alterations to parameters, such as the proportion of samples called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

3.5.2 Data Visualisation

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

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

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



(a) Customised violin plot

(b) Split violin plot

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

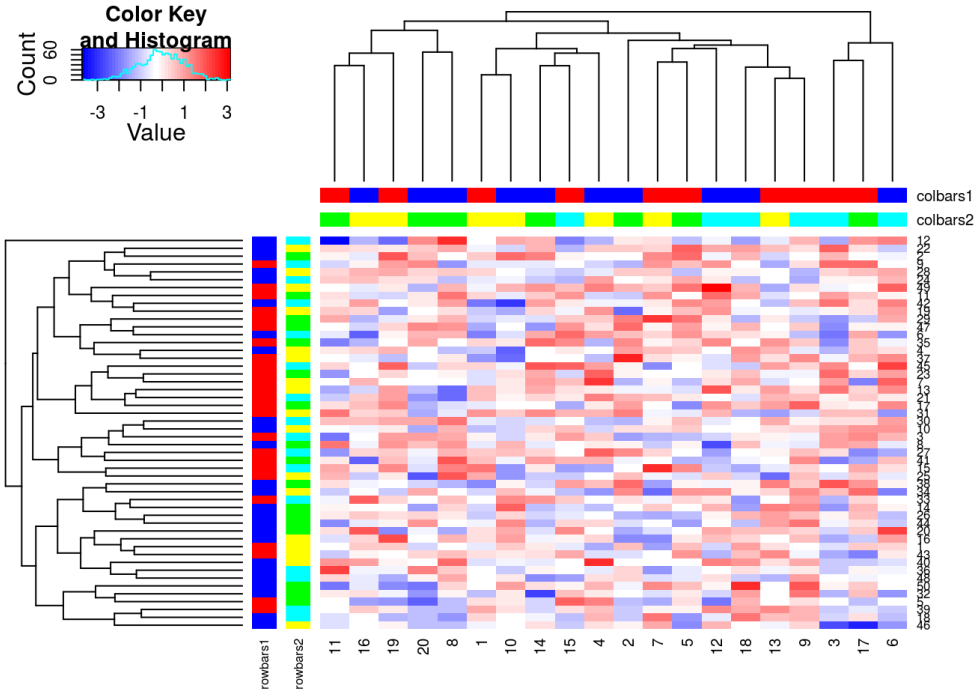


Figure 3.24: **Demonstration of annotated heatmap**. The example heatmap depicts the additional row and column annotation bars enabled by `heatmap.2x`, extending the features of `gplots` with backwards compatible inputs.

respectively. Additional parameters enable resizing of these annotation bar labels and control of reordering columns for if samples are ordered in advance (e.g., ranked by a metagene or split into groups clustered separately). These features were used through this thesis and are provided in a package GitHub repository.

3.5.3 Extensions to the iGraph Package

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

3.5.3.1 Sampling Simulated Data from Graph Structures

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

3.5.3.2 Plotting Directed Graph Structures

The `plot.igraph` package provides the `plot_directed` function specifically developed for directed graph structures and to plot activating or inhibiting for each edge (as described in section 3.4.2). As shown in Figure 3.25, this function supports separate plotting parameters for each node, node label, and edge. This includes colours of node fill, border, label text, and edges and size of nodes, edge widths, arrowhead lengths, and font size of labels. The `state` parameter for assigning activating or inhibiting to each edge determines whether edges are dipected with 30° or 90° arrowheads. Colours are assigned separately and so they may be customised. Vectorised parameters are applied

across each node or edge, whereas scalar parameters apply the same plotting parameters across them. The default layout function is `layout.fruchterman.reingold` but any layout function supported by `plot` function in `igraph` (Csardi and Nepusz, 2006) is compatible such as `layout.kamada.kawai` used to implement the Kamada–Kawai algorithm (Kamada and Kawai, 1989) for graph plots throughout this thesis.

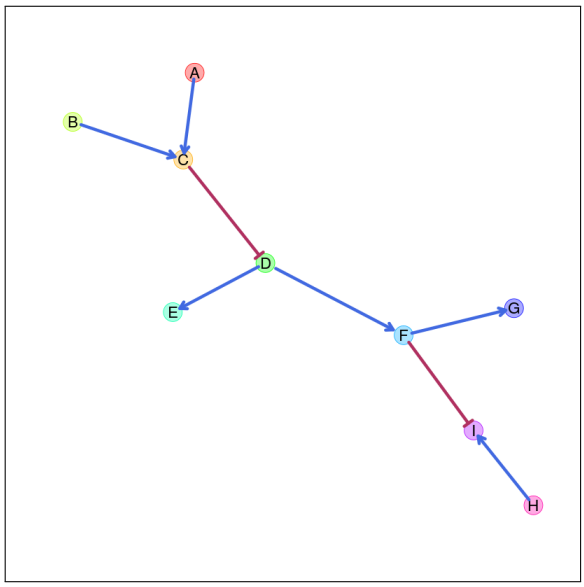


Figure 3.25: **Simulating graph structures.** An example graph structure which will be used throughout demonstrating the simulation procedure from graph structures. Here activating links are denoted by blue arrows and inhibiting links by red edges.

3.5.3.3 Computing Information Centrality

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

3.5.3.4 Testing Pathway Structure with Permutation Testing

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

3.5.3.5 Metapackage to Install iGraph Functions

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

Chapter 4

Synthetic Lethal Analysis of Gene Expression Data

4.1 Abstract

The study of networks is an interdisciplinary field which combines concepts and approaches in computer science, the fundamental principles of pure mathematics, and the applications in many fields in the social, physical, life sciences, and engineering. High-throughput technologies which gather vast amounts of molecular and cellular data have raised the need for systems-level, network-based, and genome-wide bioinformatics analysis to capture the complexity of a cell at the molecular level.

Genetic interactions (SGIs) are the deviation of a double mutant from the phenotype expected from the respective single mutants. These interactions may also occur through suppression of gene expression or protein activity by epigenetic silencing, RNA interference or drug activity. Genetic interactions have been studied at a genome-wide scale using synthetic gene array (yeast) and siRNA (nematode worm) technologies. Extension of these methods beyond model organisms is limited by the cost and labour involved and predictive models serve as an unbiased alternative to the candidate gene approach currently used in genetic screens with mammalian cell lines.

Genetic interactions have been shown to have clinical relevance with the application of BRCA1 and BRCA2 mutations with PARP1 inhibitors in Breast and Ovarian cancers, and with drug synergy between targeted therapeutics for BRAFV600E and EGFR inhibitors in Colorectal cancer. Prediction of genetic interactions has been performed in model organisms showing that protein-protein interactions, shared gene function or mutant phenotype, coexpression, or a subset of genetic interactions themselves can be

used to predict known genetic interactions.

The main focus of this research is to investigate the tissue specificity of gene networks in normal and cancerous cells. Secondary objectives include investigating the integrative network analysis of gene, protein, drug networks; the translational application of gene network analysis to predict anti-cancer drug targets and synergy; and to understand the evolutionary conservation and mechanisms underlying genetic interaction networks.

4.2 Aims and Significance

We aim to develop a network analysis approach to predict and analyse SGI networks in human cells and cancers. The main focus of this project will be to test SGIs for tissue specificity, of particular interest are SL interactions in tumours. Normal tissue will also be investigated for comparison between tissues and with tumour-specific networks. In contrast to the TCGA Pan-cancer study, which aims to find shared molecular characteristics across tumour subtypes, we aim to find molecular features (e.g., coordinated perturbation of many genes) which is unique to one or very few cell types or tumour. This is a pragmatic approach to find molecular features which, if altered, are less likely to result in adverse drug reactions.

Secondary objectives include integrative network analysis, translational research focus, and fundamental understanding of genetic interactions. Integrative analysis compares many networks across the same cell type to investigate how they are related, for instance gene regulation, genetic interaction, gene coexpression, and protein-protein interactions are known to be related. However, so far integrative analysis of networks has focused on model organisms and has not accounted for tissue specificity.

Translational research involves ensuring that the findings have clinical relevance and potential application in clinical practice. Identification of biomarkers, drug targets, and synergistic drug combinations are possible from molecular networks which can be developed for use in cancer diagnosis, prognosis, and treatment. Tools to develop predict the tissue specificity, therapeutic index, and therapeutic window of a candidate drug could be developed to prioritise RNAi and drug screens in research. If refined, these tools could be used for personalised medicine, to predict whether a particular treatment regime will be effective in a patient from their molecular profiles. Understanding the underlying mechanisms of molecular perturbations targeted for cancer treatment is important to ensure efficacy and minimal toxicity. If anticancer drugs can

be developed with drastically lower toxicity than traditional chemotherapy, they may also be feasible as a chemopreventative alternative to prophylactic surgery in high risk patients.

The fundamental understanding of genetic interactions is also important for the role of networks in heredity, cell biology, pharmacogenetics, and developmental biology. The connectivity of a network and the pathways involved in a molecular perturbation between cell types or treatment regimens can inform mechanistic molecular studies. The network properties of a cell will further enable understanding of gene expression and its role in polygenic phenotypes, complex disease, and developmental cellular differentiation. Evolutionary conservation of networks between species is important to ensure relevance of model organism studies to applications in human health and agriculture. The level of conservation between species can also be used to determine the role of gene networks in evolutionary history d identify which network features and substructures were important to be conserved across many species, and which features are unique to humans.

This project will focus on colorectal cancers which have relatively high incidence in New Zealand and are a national tissue banking initiative based in Otago. Melanomas also have high incidence in New Zealand, are a research focus at the University of Auckland, and may be useful for comparison with shared environmental and genetic risk factors (such as BRAF mutations). Breast and Stomach cancers will also be investigated to augment existing studies into synthetic lethality in the Cancer Genetics Laboratory, since CDH1 mutations are involved in both sporadic cancers and hereditary diffuse gastric cancer. Breast and Stomach cancers are cancers with some of the highest global incidence and New Zealand is no exception with typical levels for a developed country.

4.3 Background

Synthetic lethality is an emerging anti-cancer drug development approach showing promise in clinical trials as a treatment, preventative, and in combination with standard care (e.g., PARP inhibitors against BRCA mutations in Breast and Ovarian cancers). We are particularly interested in exploiting synthetic lethal interactions (where pairwise gene inactivation kills a cell) which enable development of targeted therapies so called genomic medicine informed by systems biology because they show promise as a means to effectively target tumour suppressor genes (with loss of function mutations)

to selectively kill cancerous and pre-cancerous cells. The cancer genetics laboratory are currently working on experimental screens and validation of candidate synthetic lethal partners of CDH1 (a gene implicated in hereditary and sporadic breast and stomach cancers). This project serves to develop a predictive methodology to support such experimental work with analysis of public cancer genome data to overcome many of the limitations of experimental models (namely cost, throughput, and variable genetic background). In addition to candidates based on gene expression data, we are currently extending the methodology to use DNA copy number, DNA methylation, and somatic mutation to predict synthetic lethality. Synthetic lethal predictions have the potential to scale up to genome-wide analysis enabling investigation of gene networks and tissue-specificity. Background and Methodology:

Synthetic lethality (SL) is the death of a cell or organism with the combined loss of two non-essential genes. This phenomenon was originally used to study genetic interactions and functional redundancy in models organisms (1). While synthetic lethal experiments have been performed in Drosophila melanogaster (2), Caenorhabditis elegans (3), Escherichia coli (4), Schizosaccharomyces pombe (5), and various mammalian cell lines (6), the most extensive synthetic lethal screens have been performed with the synthetic gene array (SGA) technique in Saccharomyces cerevisiae (1, 7, 8). Originally defined by double mutants, a range of mechanisms for gene inactivation of synthetic lethal partners can induce cell death including RNA interference and drug treatment where it is sometimes called induced essentiality or non-oncogene addiction in cancer research (9). Cellular viability is the main means to measure synthetic lethal effects experimentally because it is quantified and measured consistently, whereas qualitative measures of impaired organism viability are ambiguous and less relevant to yeast or cancer research.

The synthetic lethal approach to cancer therapy is a rapidly developing area of research. It has proven effective against BRCA mutations in breast cancer with the discovery of synthetic lethal interactions of BRCA1 and BRCA2 with PARP1 as distinct DNA repair functions which are mutually necessary for cellular viability (10, 11). This is particularly exciting as a proof of concept that synthetic lethality can be used to indirectly target tumour suppressor gene inactivation to selectively kill cancer cells. PARP inhibitors have been successful in numerous clinical trials in both breast and ovarian cancer against both hereditary cancers and sporadic cases of BRCA mutant cancer (12). Not only do synthetic lethal drugs have the potential to be effective across multiple cancer types, they have could be utilised for chemopreventation against hered-

itary cancers in high risk individuals with the ability to achieve high therapeutic index with this approach (6, 13). Synthetic lethality has also been explored as a means to target oncogenes which are difficult to selectively target directly due to high sequence homology to their wild-type counterpart or other genes (14).

The cancer genetics laboratory are currently working on developing a synthetic lethal approach to target the tumour suppressor gene CDH1 which has been found to cause predispose early-onset breast and stomach cancers in mutation carriers, including families of New Zealand Māori (15, 16). These families are currently closely monitored and offered drastic preventative surgery. If it were developed drug selective against CDH1 mutant tumours would serve not only as a chemopreventative alternative for these families but also benefit the wider community as a treatment for sporadic cases of CDH1 mutant cancer. To augment experimental work on CDH1 with isogenic cell lines, a computational methodology is explored here to exploit public cancer genomic databases.

Figure 1. Impact of various negative (a) and positive (b, c) synthetic genetic interactions on growth viability fitness in yeast. Adapted from Costanzo, Baryshnikova (8).

Microarray and massively-parallel sequencing technologies are driving a revolution in molecular biological research, particularly with regard to cancer where the premise of genomic medicine is rapidly becoming feasible with the use of genomics to identify cancer genes, diagnose patients with actionable mutations, and use gene expression as a prognostic marker. Genomic data could also be used to identify novel drug targets and synthetic lethal partners of known cancer genes in particular. The Cancer Genome Atlas database (TCGA) and the overarching International Cancer Genome Consortium (ICGC) provide a valuable public cancer genome data resource because they support many different data types for the same samples, for many different cancer types, and for high sample sizes (17-19). They host data of patient clinical factors, gene expression, somatic mutation, DNA copy number, and DNA methylation which could all serve to predict synthetic lethality from frequency of mutually exclusive gene inactivation and its impact on patient survival. A number of other databases are given in the Table 6 which may be used to explore gene function, drug target feasibility, or replicate analyses but TCGA and ICGC datasets will be the focus of this project.

Networks are an established research area of pure mathematics producing many applications relevant to biology including evolutionary trees, metabolic pathways, gene regulation, and protein-protein interactions (20). It is a branch of graph theory which

deals with the connections between discrete objects which includes terminology for particular interaction patterns, visualisation methods, and algorithms to predict and measure individual interactions and the whole network (21). This established set of mathematical tools could be utilised to use a systems biology approach to using genomic data to predict synthetic lethal interactions or analyse patterns in the resulting predictions or supporting experimental data.

Network medicine is an emerging notion that network analysis of biology could be useful for clinical applications and translational research including identification of disease genes (for diagnostics), biomarkers (for prognosis), identify novel drug targets, and find the biological significance of mutations and SNPs found by genome-wide association and whole-genome sequencing approaches (22). Molecular networks may be useful to understand perturbations of cellular functions in human disease including groups of genes underlying multiple separate or co-occurring diseases and whether they occur in a tissue specific manner. A network understanding of a disease may be relevant not only to genetic risk and mutation but also to the impact of the disease (or causes of it) in abnormalities in metabolic pathways, protein complexes, epigenetic marks, and microRNAs. An understanding of cellular function is important for network pharmacology, a modern approach to drug design where understanding of the effect of the drug on the network is more important than specificity to a single target (23). Combined or synergistic drug targets are a known means to more effectively treat diseases, exploiting the network enables targeting disease genes indirectly (including synthetic lethal partners) and use of drugs with multiple targets (known as polypharmacology).

There is a growing need for a robust approach to cost-effective prediction of candidate synthetic lethal interaction, particularly in cancer research. Exploiting existing public genomic databases is an ideal way to utilise existing resources with suitable sample sizes, data types, and different limitations to those of laboratory experiments. A number of computation approaches to synthetic lethality have been developed but many of these rely on data not available to cancer researchers, methods that are difficult to replicate, overfitted to a particular dataset, having mixed validation results, or do not have a software tool accessible to the research community. These methodologies will still be considered to develop an improved synthetic lethal interaction prediction tool (SLIPT). The methodologies summarised below in Table 1 include those reviewed by Van Steen (24) or Boucher and Jenna (25).

Table 1. Existing prediction methods for Genetic Interaction Networks Therefore the data types considered to be predictive of synthetic lethality and the biological

questions that could be addressed by them are summarised in the Figure 2.

Figure 2. Mindmap of synthetic lethal predictors and biological areas of relevance. Underlined points have been investigated with preliminary data, italicised points are being considered in the immediate future of this project.

A bioinformatics approach has distinct limitations to experimental methods and would work well combined with genetic screen data and conventional molecular biology laboratory validation techniques to answer biological research questions. Compared with an experimental screen, a bioinformatics approach has the benefits of reduced costs, with the potential for automation, scaling up, and replication of the same gene across populations and cell types. Analysis of public genomic data accounts for real tumour variation showing detection with tumour heterogeneity and genomic instability. Compared with a cell line or xenograft experimental model we are limited by difficulties in establishing validity of a novel method, lack of mechanism, or potential for testing drug activity in the same system. This method may further miss useful therapeutic candidates from variable genetic background and be limited by the population sampled.

It is notable that another group has recently published a methodology with a similar purpose which they have called DAISY (Data Mining Synthetic Lethality Pipeline). Their methodology covers some of research objectives initially planned for this project, however, many findings are yet to be met by the existing methodology and Jerby-Arnon, Pfetzer (41) have yet to provide a means for researchers to replicate their method or a software package to apply it to new datasets. Some of the findings of Jerby-Arnon, Pfetzer (41) are helpful such as the observation that DNA copy number is comparable in power to detect synthetic lethality as gene expression with their methodology. We had not considered this ourselves since our gene of interest, CDH1, is not widely variable in copy number in tumours. This lead to testing whether the current SLIPT methodology could be adapted to work with DNA copy number data and further investigation into whether somatic mutation or DNA methylation could be similarly utilised. Jerby-Arnon, Pfetzer (41) showed not only that publicly available tumour data was able to predict enrichment of shRNA synthetic lethal screen hits for VHL in renal cell lines but also that bioinformatics analysis of cell line data was similarly applicable.

This project builds upon prior work during my study towards Honours in genetics which involved developing a synthetic lethal interaction prediction tool (SLIPT) from public gene expression microarray data (43). This methodology compares the distribution of samples for pairs of genes with the premise that synthetic lethality would lead to a deficit of samples showing inactivity of both genes. A chi-squared test of quantiles

was used to assess significance of the interaction along with a directional criteria as shown in Figure 3. This methodology has been adapted and executed for analysis of RNA-Seq expression data, DNA copy number from SNP microarrays, to run in parallel on high performance computing resources, and for tumour suppressor (TS_SL) or oncogenes (Onco_SL analogous to DAISY predicting synthetic dosage lethality; SDL).

Figure 3. Schematic outline of bioinformatic synthetic lethal prediction approach.

4.4 Background

Synthetic lethality (SL) is the death of a cell or organism with the combined loss of two non-essential genes. This phenomenon was originally used to study genetic interactions and functional redundancy in models organisms (Boone et al. 2007). While synthetic lethal experiments have been performed in Drosophila melanogaster (Dobzhansky 1946), Caenorhabditis elegans (Lehner et al. 2006), Escherichia coli (Butland et al. 2008), Schizosaccharomyces pombe (Roguev et al. 2007), and various mammalian cell lines (Kaelin 2005), the most extensive synthetic lethal screens have been performed with the synthetic gene array (SGA) technique in Saccharomyces cerevisiae (Boone et al. 2007; Costanzo et al. 2011; Tong et al. 2004).

Originally defined by double mutants, a range of mechanisms for gene inactivation of synthetic lethal partners can induce cell death including RNA interference and drug treatment where it is sometimes called induced essentiality or non-oncogene addiction in cancer research (Fece de la Cruz et al. 2015). Cellular viability is the main means to measure synthetic lethal effects experimentally because it is quantified and measured consistently (as shown in Figure 1), whereas qualitative measures of impaired organism viability are ambiguous and less relevant to yeast or cancer research.

The cancer genetics laboratory are currently working on developing a synthetic lethal approach to target the tumour suppressor gene CDH1 which has been found to cause predispose early-onset breast and stomkach cancers in mutation carriers, including families of New Zealand Māori (Berx et al. 1995; Guilford et al. 1998). These families are currently closely monitored and offered drastic preventative surgery. If it were developed, a drug selective against CDH1 mutant tumours would serve not only as a chemopreventative alternative for these families but also benefit the wider community as a treatment for sporadic cases of CDH1 mutant cancer. To augment experimental work on CDH1 with isogenic cell lines (Telford et al. 2015), a computational methodology is explored here to exploit public cancer genomic databases.

There is a growing need for a robust approach to low-cost prediction of candidate synthetic lethal interaction, particularly in cancer research. Exploiting existing public genomic databases is an ideal way to utilise existing resources with suitable sample sizes, data types, and different limitations to those of laboratory experiments. A number of computation approaches to synthetic lethality have been developed but many of these rely on data not available to cancer researchers, methods that are difficult to replicate, over-fitted to a particular dataset, having mixed validation results, or do not have a software tool accessible to the research community. These methodologies have been reviewed in a literature review to inform the development of a synthetic lethal interaction prediction tool (SLIPT) using gene expression or mutation data (as shown in Figure 2) from the Cancer Genome Atlas Project (TCGA) and to inform interpretation of the results.

Figure 1. Impact of various negative synthetic genetic interactions on growth viability fitness in yeast. Adapted from Costanzo et al. (2011).

Figure 2. Schematic outline of bioinformatic approach to synthetic lethal prediction for partners of a query gene with Chi-Square test, directional condition and adjusting for multiple tests.

4.5 Sourcing TCGA data

4.6 Quality checking

4.7 Global Synthetic Lethality

Global levels of synthetic lethality were analysed as part of my Honours project to address concerns of high numbers of synthetic lethal candidates for CDH1. This turned out to be typical for most genes in the microarray dataset. Due to newer samples and concerns about sample quality in TCGA microarrays, RNA-Seq datasets were used here. As my PhD will focus on RNA-Seq data for gene expression, this was replicated using the TCGA breast cancer RNA-Seq dataset on the New Zealand eScience Infrastructure Intel Pan supercomputer.

4.8 CDH1 Analysis with Subgroups

As discussed previously, CDH1 (also known as E-Cadherin) is a tumour suppressor gene and the subject of ongoing investigations in the Cancer Genetics Laboratory. Synthetic lethal gene candidates for CDH1 from RNA-Seq expression data have been the subject of most of my PhD beginning with replication of previous pathway over-representation analyses in RNA-Seq data (Araki et al. 2012). A novel finding compared to previous analyses in microarray data was correlation structure in the expression of candidates synthetic lethal genes in CDH1 low tumours (lowest 1/3rd quantile of expression). Subgroups of genes were enriched for distinct biological pathways and elevated in different clusters of samples including some by clinical factors such as estrogen receptor status.

These results were presented in a poster at the QMB Cancer and Drugs satellite meeting in 2014. More recent analyses have also investigated intrinsic (PAM50) subtype and somatic mutation (of highest impact genes) against these gene clusters.

4.9 Cell Line Analysis

As breast cancer cell lines are the experimental system in which many cancer genetics and drug targets are investigated, these were analysed in addition to patient samples from TCGA. The cancer cell line encyclopaedia (CCLE) is a resource for genomics profiles across a range of cell lines. These have also been used to generate synthetic lethal candidates for comparison to those in experimental screen and predictions from TCGA expression data. A transcriptome experiment has been conducted by the Cancer Genetics Laboratory to test their CDH1-/- null MCF10A cell lines compared to an otherwise isogenic wildtype (Chen et al. 2014). While differential expression analysis was inconclusive due to few technical replicates, this data was also useful to determine genes which were not detectable in MCF10A cell lines which would not be expected to detect synthetic lethality in siRNA screen data even if they were predicted to be synthetic lethal in expression data.

4.10 Mutation, Copy Number, and Methylation

Due to promising synthetic lethal data on mutation and DNA copy number analyses (Jerby-Arnon et al. 2014; Lu et al. 2015), these were also investigated to compare genes for synthetic lethality in an analogous manner to expression analyses in the TCGA data.

Due to the low somatic mutation rate (and lack of available) germline mutations for many genes, it was not possible to detect many double mutations with significantly under-representation in cancers. There were also concerns about using rare mutations with unknown significance or excluding functional mutations by only using those in the exons. It was possible to compare deletion and duplication of DNA copy number in a manner analogous to expression quantiles. However, these overlapped poorly with candidate interacting partners from expression analyses and concerns were raised that they may not be relevant to CDH1 which is typically inactivated in tumours by loss of function mutations or DNA methylation (PJ Guilford, personal communication).

DNA methylation data was also prepared for synthetic lethal analysis but was discontinued due to computational challenges, expected similarity to expression results, difficulty defining loss of function methylation at a gene level across CpG sites, and the concerns raised in the next section.

4.11 ANOVA of Expression Predictors

Another approach was to only use copy number, mutation, or hyper-methylation data for genes in which they would impact on gene function and occur frequently in tumours. Before investigating whether these impact on gene function, they were investigated as predictors of variation in gene expression. If these are not giving variation independent of gene expression, expression would be a more suitable measure of gene function as it is widely generated in studies and useful as a clinical biomarker.

Globally predicting gene expression across all genes from DNA copy number and somatic mutation was attempted by ANOVA. However, this was computationally challenging and gene-specific analyses would be more informative. Gene specific ANOVA and linear regression was performed but was raised more issues than it addressed. There were issues with interaction terms and mutation data, many genes were not tested for these since there were so few mutations for these genes in the dataset. It was possible to include DNA methylation in gene-specific analyses (despite the concerns raised above) but the R2 values for each gene were still generally very low and issues with insufficient mutant samples for interaction terms became worse. This means that the approach used differs for each gene making it difficult to compare them. The challenges raised here suggested that expression is very difficult to predict with other factors but including these other factors would be difficult and plagued by multiple-testing, particularly comparing between them with the current synthetic lethal prediction method. This

led to investigations into the simulation of synthetic lethality.

4.12 Mutation Analysis, Pathway Expression, and Metagene Synthetic Lethality

Pathway data was sourced from a variety of databases including the Kyoto Encyclopaedia of genes and genomes (KEGG), Gene Ontology, and Reactome using their R packages and WikiPathways (by parsing gpml files). These were used to analyse generate Pathway-based gene sets for expression clustering and synthetic lethal analysis. The focus of later analyses are the Reactome results because of their concordance with experimental results (in preliminary analyses), containing a large portion of the genome while being recently updated and curated based on the literature. Gene Signatures were also sourced from (Gatza et al. 2011; Gatza et al. 2014) to check effects known to occur in breast cancer, such as upregulated pathways in particular subtypes behave as expected.

Metagenes (from the first eigenvector of the singular value decomposition) do not necessarily follow the direction of activation of the pathway. Metagenes were multiplied by -1 if they did not positively correlate with the mean centroid of the pathway across samples so the metagenes were in the direction of the majority of genes while preserving the metagene weighting. This assumes that most genes are involved in the activation of the pathway while auxiliary regulators and inhibitors are the minority. The metagenes for the gene signatures were in the direction expected reassuring concerns that direction of metagenes would affect synthetic lethal prediction. Therefore pathway metagenes could be used to predict synthetic lethal pathways using reactome pathways against CDH1.

The metagenes were also used to heatmap pathways and gene signatures across the samples to compare against clinical factors as performed with genes. As alluded to earlier, somatic mutation for the genes with the highest predicted impact and frequency were also added to both the pathway and gene heatmaps. However most mutations were inconclusive apart from p53 which was over-represented in estrogen receptor negative tumours and under-represented in CDH1 deficient tumours. The main groups with drastically different gene expression profiles are estrogen receptor negative tumours and normal samples, both of which have been excluded from some analyses in an attempt to find subtype specific effects. While estrogen receptor positive and negative tumours have distinct synthetic lethal genes and pathways, these have not been investigated

in detail and remain to be revisited once a pathway analysis method has been settled upon.

Another use of mutation data was to investigate gene expression in CDH1 mutant and wildtype samples (as defined by non-synonymous somatic mutation). Differential expression and synthetic lethal analysis have both been performed using this mutation data in addition to the prior CDH1 low analysis using solely gene expression data.

4.13 Data clean up, gene SL, and pathway SL

Due to concerns about the quality of TCGA data, the latest version of the TCGA breast cancer data from the ICGC data portal in August 2015. This added several hundred sample not contained in previous analysis (up to n=1177). However, clustering analysis of the correlation matrix found a number of samples with poor correlation to the rest of the group. These either had no read count or stem from the same source site or a patient with a rare metaplastic subtype. This suggested issues with sample quality or laboratory handling, many of these sample were done in triplicate (rare from this dataset) and correlated poorly with technical replicate samples. Therefore a final dataset of 1168 samples (112 normal, 1049 primary tumour, and 7 metastasis) were used to repeat many of the above analyses. Clinical and mutation data was also updated for this new analysis including adapting the PAM50 subtyping method (Parker et al. 2009), established for microarrays, to RNA-Seq using the new training centroids on RSEM normalised data (JS Parker, personal communication).

While results presented at this meeting may resemble previous results, they are all based on an entirely new TCGA expression data set using voom normalisation (Smyth 2005) on the raw read counts data, rather than RSEM provided in TCGA tier 3 (apart from the aforementioned PAM50 subtyping). Synthetic lethal analysis, clustering, and heatmaps have been re-done on this new dataset for synthetic lethal genes and pathways against both CDH1 expression and mutation in all samples, tumour-only, and Estrogen receptor specific. This has generated, not only synthetic lethal gene candidates and those overlapping with siRNA screen candidates but also their over-represented pathways (from genesetDB) and synthetic lethal candidate pathway metagenes. This reproduces some results consistent with experiments of the Cancer Genetics Lab, including a role of GPCR pathways. Also notable are some of the pathways which were not detected in the siRNA screen, including immune signals which we would not expect in isolate cells but are still known to be involved in recent cancer treatment strategies

(Olszanski 2014).

4.14 Overview of Challenges

Previous gene expression analysis and comparisons to experimental screen data (Telford et al., 2015) led to some interesting synthetic lethal candidate genes for CDH1 and enriched pathways in subgroups. Of particular interest was enrichment of G Protein Coupled Receptors and related pathways in some subgroups supporting the hypotheses and experimental results with the MCF10A cell line performed by the Cancer Genetics Laboratory. It has also been noticed that some of the other candidate biological pathways such as immune functions are known to have important roles in breast cancers but would not be detected in the cell line experiments in isolated culture.

However, there remain concerns about the underwhelming overlap between bioinformatics predictions and cell line experiments and inconsistent gene candidates across datasets or analyses. Simulation analysis and multiple testing have also raised statistical concerns, particularly at the gene level. Hence, the current focus of this project is to identify biological pathways with evidence of synthetic interactions in E-Cadherin (CDH1) deficient breast cancers. Biological pathways present fewer issues with multiple testing and are synthetic lethal pathways are known to be conserved more between species than synthetic lethal genes (Dixon et al. 2008).

There are several approaches to synthetic lethal pathway analysis, I will present results for both gene set over-representation analysis (using GeneSetDB) from predicted synthetic lethal gene partners and synthetic lethal predictions using metagenes (generated by the singular value decomposition). Both of these analyses use the Reactome database to define pathways, this database also has pathway structure data which has also been used to construct a network and perform information centrality analysis as a measure of gene essentiality. These analyses use an updated dataset of 1168 TCGA Breast samples with samples removed using the voom normalised raw count data and due to data quality concerns raised by other students working with TCGA data in our research group. Synthetic lethality has been test against both low CDH1 expression (exprSL) and non-synonymous CDH1 (somatic) mutation (mtSL) in many of these analyses. While low expression is a promising biomarker and a proxy for reduced gene activity (whereas there remain questions around whether mutations are functional, detectable, or expressed), however, mutations have also been considered since null mutations were used for an experimental model and they are relevant to

HDGC patients.

The overlap between synthetic lethal from bioinformatics SLIPT predictions and siRNA screening has raised other questions including whether the number of genes and pathways enriched would be expected by chance. This of particular concern since the siRNA candidate genes themselves are highly enriched for particular pathways so selecting any intersect with them would be enriched for these pathways. The siRNA data is also based on cell line models which have limitations in application to a genetically variable patient population with a complex tumour microenvironment interacting with immune cells. One approach is to compare the candidate genes is to exclude genes that were not tested in both systems, such as those not expressed in cell lines or those with more than 1/3 of TCGA patients without any RNA Seq reads so the lowest quantile cannot be defined for SLIPT analysis. Another approach is to test whether pathways are enriched in randomly sampled genes, comparing many resampled or permutations of these genes to the enrichment statistics observed for these pathways in the SLIPT candidates and their intersection with the siRNA hits shows whether we detect these pathways more than we expect by chance. Both of these are being applied with developing a method and overcoming technical challenges for the latter being the focus of recent work. The main challenge at the moment is to compare SLIPT results to experimental candidates and explain why so few genes (and so many pathways) overlap.

4.15 Comparison of gene SL predictions and siRNA screen candidates

As discussed above, comparing genes between experimental screen candidates and prediction from TCGA expression data has been difficult. Figure 3 summarises the approaches to comparing genes accounting for some of the differences between the datasets. Of particular concern are the over-represented pathways in genes detected by both methods. There is no statistical evidence that SLIPT predicted genes or siRNA candidates are enriched in with each other. The siRNA candidates themselves are over-represented with many pathways including GPCRs so any intersection with these would contain some of these pathways. Whether these pathways are contained in the intersection more than expected by chance is the problem the two approaches below were designed to tackle.

Figure 3. A summary of the challenges and approaches involved in the comparison of synthetic lethal candidates from bioinformatics analyses to siRNA experimental

screen data.

4.16 Permutation or Re-Sampling of genes for pathway enrichment.

Approach 1: assumes that the size of the intersection is fixed at the observed size of 450 or 335 for exprSL and mtSL respectively. A random sample of this size is taken and tested for enrichment of all 1652 Reactome pathways. This is added to a random sample of the remaining genes of the observed size 3576 or 2233 for exprSL and mt SL respectively and tested again for enrichment of all Reactome pathways. This was repeated 10,000 times on the New Zealand eScience Infrastructure Intel Pan supercomputer to generate a null distribution of expected Chi-Squared values for each pathway to compare to the SLIPT predictions and the intersection with experimental screen genes. Empirical p-values were defined by the proportion of the 10,000 null Chi-Squared values which were greater or equal than the observed before being adjusted for multiple tests by the number of Reactome pathways.

Approach 2: assumes that the size of the intersection is varies and tests whether it is significantly enriched or depleted for siRNA genes. A random sample of the observed size for predicted genes of total 4026 or 2568 for exprSL and mtSL respectively is taken and tested for enrichment of all 1652 Reactome pathways. This is also used to derive an intersection with siRNA screen candidates and is tested again for enrichment of all Reactome pathways. This was repeated 10,000 times on the New Zealand eScience Infrastructure Intel Pan supercomputer to generate a null distribution of expected Chi-Squared values for each pathway to compare to the SLIPT predictions and the intersection with experimental screen genes. Empirical p-values were defined by the proportion of the 10,000 null Chi-Squared values which were greater or equal than the observed before being adjusted for multiple tests by the number of Reactome pathways. The size of each sampled intersection was also used to show that more than 5% of samples contained an intersection lesser and greater than the number of genes. So despite concerns about the number of genes detected, there was no evidence of less genes than expected by chance either, the composition of genes may still yield candidates.

4.17 Comparison of candidate SL Pathways

Thus we have identified candidate synthetic lethal pathways by gene set over-representation, metagene synthetic lethality, and re-sampled empirical pathway over-representation. The challenge currently under consideration is whether these methods can be compared and which may lead to biologically meaningful or clinically relevant synthetic lethal candidate pathways.

4.18 Future Directions

As discussed before, there are a number of future directions within the scope of this project. A number are being considered including revisiting simulations to include pathway structure, network-based analyses, and continue investigations into synthetic lethal genes and pathways in clinical subgroups. A number of these are given in as an example in the following Timeline. The synthetic lethal analysis to generate candidate pathways for other genes and in other cancer datasets is another avenue which has been left as an opportunity for a new student since repetition of these methods would not develop more skills or demonstrate the critical understanding of the field. There are a numerous experimental and clinical challenges involved in seeing any synthetic lethal candidates into preclinical models, clinical trials, or understanding the functional and mechanistic basis and implications of these interactions. These approaches are better suited to researchers with different skills as those involved in ongoing synthetic lethal experiments in the Cancer Genetics Laboratory.

4.19 Hub Genes

4.20 Metagene pathway expression

4.21 Metagene synthetic lethality

4.22 Replication in stomach cancer

4.23 Important Results

Table 1. Hub gene function in TCGA breast cancer microarray expression SL predictions (n=600).

Table 2 Hub gene function in TCGA breast cancer RNA-Seq expression SL predictions (n=878).

Table 3. Hub gene function in BC2116 breast cancer microarray expression SL predictions (n=2116).

Figure 3. Heatmap of RNASeq gene expression in predicted SL partners of CDH1 showing distinct subgroups of SL partners and links between SL partner expression and clinical variables.

Table 5. Gene set enrichment results for subgroups of CDH1 SL partners shows functional variation.

As discussed in the previous committee meeting, we have developed a simple, interpretable, computational approach to predict synthetic lethal partners from genomics data. Originally developed for microarray gene expression data, it has been expanded to test DNA copy number, or RNA-Seq gene expression data which are both also supported by the TCGA dataset. DNA copy number was included for comparison with the DAISY tool of Jerby-Arnon et al. (2014). Predictions based on microarray data were inconclusive when compared with an RNAi screen for CDH1 in MCF10A breast cells as performed by Telford et al. (2015), few predictions replicated between BC2116, CCLE, or TCGA microarray datasets, results with gene expression and DNA copy number were vastly different, and predictions from TCGA microarray and RNA-Seq datasets for the same samples differed were inconsistent. The Aligent TCGA microarray data in particular is difficult to compare to other datasets and will in the future use Affymetrix microarrays or RNA-Seq platforms for predictions from gene expression

data. The analyses focus on gene expression data as it is widely available for applications in other cancers and current attempts to use gene expression data for synthetic lethal discovery vary widely (Jerby-Arnon et al. 2014; Lu et al. 2015; Tiong et al. 2014). There is no consensus for which approach is more appropriate since they lack much a basis on biological experimental data or statistical modelling and often use difficult to interpret machine learning methodology.

Genomics analyses are prone to false-positives and require statistical caution, particularly where working with gene-pairs scale up the number of multiple tests drastically, at the expense of statistical power. Experimental SGA and RNAi screens for synthetic lethality are also error-prone, especially with false-positives, raising the need for understanding the expected behaviour and number of functional relationships and genetic interactions in the genome, or in discovery of synthetic lethal partners of a particular query gene. A characteristic of gene interaction networks is a scale-free topology leading to highly interacting hub genes, these represent important genes in a functional network. As shown in Tables 1-3, Gene Ontology terms for genes important in cancer proliferation, progression, and drug response were enriched in hub genes, showing that synthetic lethal interactions are among important genes in cancer cells. Gene functions replicated across the breast cancer datasets are highlighted in bold, despite differences in particular hits, gene expression platforms, and only correcting for multiple tests for each gene query separately, there are many gene functions replicated across breast cancer gene expression analyses. TCGA microarray data was less consistent with the other datasets, as expected from lower sample size, lower concordance of particular hits for the example query of CDH1, and suspected lower quality of data on the Aligent microarray platform.

As specific genes were difficult to replicate across experiments, gene expression profiles for synthetic lethal partners must be more complex than originally expected to directly compensate for loss of query gene or completely lack (or clearly under-represent) co-loss (Jerby-Arnon et al. 2014; Kelly 2013; Lu et al. 2015). The predicted synthetic lethal partners of CDH1 (with FDR correction) were investigated with gene expression profiles and clinical variables to find relationships in gene expression, gene function, and clinical characteristics. The large number of hits indicate that synthetic lethality is error-prone and identifying genes or pathways relevant for clinical application will be difficult.

The expression profiles of the SL partners of CDH1 predicted from the TCGA breast cancer RNA-Seq data in CDH1 low tumours (where synthetic lethal partners

are expected to have compensating high or stable expression) are shown in Figure 7 and their corresponding functional enrichment is given below in Table 5, computed as WikiPathways in GeneSetDB (Araki et al. 2012). The 3 subgroups of genes are showed functional organisation of expression profiles in CDH1 low breast tumours. The first group is enriched for G protein coupled receptors, an established drug target and supported in cell line experiments (Telford et al. 2015). The second group contains genes involved in development and metabolism consistent with cancer cells showing stem cell properties and the Warburg hypothesis (Merlos-Suarez et al. 2011; Warburg 1956). The third group contains cell signalling and focal adhesion functions, including pathways involved in cancer proliferation, metastasis, and consistent with internal synthetic lethality within the pathways containing CDH1 (Barabsi & Oltvai 2004).

Ductal breast cancers show higher expression of synthetic lethal partners suggesting treatment would be more effective in this tumour subtype. However, there is consistently low expression of SL partners in ER negative tumours, although this is independent of tumour stage and consistent with poor prognosis in these patients and could inform other treatment strategies or prevent ineffective treatment further impacting quality of life in these patients. These results suggest that synthetic lethal partner expression varies between patients; that these different tumour classes would react differently to the same treatment; that treatment of different pathways and combinations in different patients is the most effective approach to target genes compensating for CDH1 gene loss; and the expression of synthetic partners could be a clinically important biomarker. While these are important clinical implications, the synthetic lethal predictions lack enough confidence for direct translation into pre-clinical models or clinical applications leading to a need for statistical modelling and simulation of synthetic lethality in genomics expression data.

Chapter 5

Pathway Structure of Synthetic Lethal Genes

5.1 Abstract

Effective screening, prediction, and analysis of synthetic lethal interactions are a crucial part of developing next generation anti-cancer strategies. Therefore, we propose developing a computational statistical procedure to identify synthetic lethal interactions and construct gene networks. This will enable the development of personalised medicine targeted to particular molecular aberrations. Genetic tests and genomics have the potential to revolutionise cancer screening, diagnosis, and prognostics; targeted therapeutics, similarly, have applications in prevention and therapy of sporadic or hereditary cancers with known molecular properties.

Construction of genetic interaction networks is important to understand the functional complexity of cellular and molecular biology, particular how it relates to existing networks for protein binding, gene regulation, genetic interaction, and gene co-expression. Comparison of networks between species will enable use of known interactions in yeast and understanding of the evolutionary importance of genetic interactions. Comparing protein and gene networks is valuable to determine which are more effective for prediction of drug targets and development of biomarkers.

Comparison of networks between cells could lead to clinically significant findings and development of effective anti-cancer drugs: both comparison between normal and cancerous cells from the same tissue and comparison across tissues. Among the most exciting applications is the use to prioritise drug screening and repurpose existing drugs. Genetic interaction discovery and gene network analysis also have the poten-

tial to develop predictions for a drug's therapeutic index, tumour specificity, tissue independence, and synergistic interactions based on known targets.

5.2 Background

5.3 Reactome Network structure and Information Centrality as a measure of gene essentiality

Network structure is another useful strategy to analyse gene function and this has been used to investigate network properties of a network constructed from of Reactome pathways imported with the paxtoolsr R package (Demir et al. 2010). Most notably, information centrality which has been proposed as a measure of gene essentiality was calculated as performed by Kranthi et al. (2013) using the efficiency and shortest path between each pair or nodes in the network before and after a node of interest is removed to test the importance of a node to network connectivity. Reactome contains substrates and cofactors in addition to genes or proteins, supporting the idea of centrality as a measure of essentiality, a number nodes with the highest centrality were essential nutrients including $Mg2^{+}$, $Ca2^{+}$, $Zn2^{+}$, and $Fe3^{+}$.

5.4 Synthetic lethal genes in synthetic lethal pathways

5.5 Methods

5.5.1 Sourcing graph structure data

5.5.2 Constructing pathway subgraphs

5.5.3 Centrality Measures

5.5.4 upstream and downstream gene detection

5.5.5 permutation analysis

5.6 Centrality and connectivity of synthetic lethal genes

5.7 Upstream or downstream synthetic lethal candidates

5.8 Hierachical approach

5.9 Discussion

5.10 Conclusion

Chapter 6

Simulation and Modeling of Synthetic Lethal Pathways

6.1 Abstract

Synthetic lethal interactions occur between genes when their combined loss is deleterious to a cell due to their shard essential function. These interactions are the basis of emerging anti-cancer drug design strategies against specific loss of genes in cancers, such as CDH1 in breast cancers. As discussed in previous meetings, we have developed a bioinformatics gene-expression analysis approach complementary to high-throughput RNAi screening in pre-clinical models. This approach successfully scaled up computationally, adapted to a range of microarray and RNA-Seq datasets, and applied to DNA copy number and somatic mutation data. However, there are difficulties replicating between datasets and RNAi candidates, such as the synthetic lethal screen for CDH1 partners in MCF10A breast cells (Telford et al. 2015). It is unclear whether this stems from different sources of error between methodologies, tissue specificity of synthetic lethal interactions, or the approaches detect different classes of genetic interactions. Therefore, we construct a statistical model of synthetic lethality to understand the sources of error in our approach and analyse simulated data to test how many synthetic lethal partner genes can be detected from gene expression data. We have developed a model using multivariate normal distributions of expression levels to test the effects of correlation structure, number of genes, number of samples, and underlying number of true synthetic lethal partners on the error rate and distribution of chi-squared test statistics. There is structural and functional complex in gene expression profiles of predicted CDH1 synthetic lethal partners. We intend to develop correlation structure

simulations to model biological pathways and comparing simulations with real gene expression data. Analysis and prediction of gene networks, feasible or robust drug targets, and biomarkers of drug response are further directions for this project.

6.2 Background

Synthetic lethality (SL) is the death of a cell or organism with the combined loss of two non-essential genes. This phenomenon was originally used to study genetic interactions and functional redundancy in models organisms (Boone et al. 2007). While synthetic lethal experiments have been performed in Drosophila melanogaster (Dobzhansky 1946), Caenorhabditis elegans (Lehner et al. 2006), Escherichia coli (Butland et al. 2008), Schizosaccharomyces pombe (Roguev et al. 2007), and various mammalian cell lines (Kaelin 2005), the most extensive synthetic lethal screens have been performed with the synthetic gene array (SGA) technique in Saccharomyces cerevisiae (Boone et al. 2007; Costanzo et al. 2011; Tong et al. 2004).

Originally defined by double mutants, a range of mechanisms for gene inactivation of synthetic lethal partners can induce cell death including RNA interference and drug treatment where it is sometimes called induced essentiality or non-oncogene addiction in cancer research (Fece de la Cruz et al. 2015). Cellular viability is the main means to measure synthetic lethal effects experimentally because it is quantified and measured consistently (as shown in Figure 1), whereas qualitative measures of impaired organism viability are ambiguous and less relevant to yeast or cancer research.

The cancer genetics laboratory are currently working on developing a synthetic lethal approach to target the tumour suppressor gene CDH1 which has been found to cause predispose early-onset breast and stomach cancers in mutation carriers, including families of New Zealand Māori (Berx et al. 1995; Guilford et al. 1998). These families are currently closely monitored and offered drastic preventative surgery. If it were developed, a drug selective against CDH1 mutant tumours would serve not only as a chemopreventative alternative for these families but also benefit the wider community as a treatment for sporadic cases of CDH1 mutant cancer. To augment experimental work on CDH1 with isogenic cell lines (Telford et al. 2015), a computational methodology is explored here to exploit public cancer genomic databases.

Microarray and massively-parallel sequencing technologies are driving a revolution in molecular biological research, particularly with regard to cancer where the premise of genomic medicine is rapidly becoming feasible with the use of genomics to identify

cancer genes, diagnose patients with actionable mutations, and use gene expression as a prognostic marker. Genomic data could also be used to identify novel drug targets and synthetic lethal partners of known cancer genes in particular. The Cancer Genome Atlas database (TCGA) and the overarching International Cancer Genome Consortium (ICGC) provide a valuable public cancer genome data resource because they support many different data types for the same samples, for many different cancer types, and for high sample sizes (Cancer Genome Atlas Research Network 2014; Cancer Genome Atlas Research Network et al. 2013; International Cancer Genome Consortium 2014). They host data of patient clinical factors, gene expression, somatic mutation, DNA copy number, and DNA methylation which could all serve to predict synthetic lethality from frequency of mutually exclusive gene inactivation and its impact on patient survival. A number of other databases are given in the Table 6 which may be used to explore gene function, drug target feasibility, or replicate analyses but TCGA and ICGC datasets will be the focus of this project.

Figure 1. Impact of various negative (a) and positive (b, c) synthetic genetic interactions on growth viability fitness in yeast. Adapted from Costanzo et al. (2011).

There is a growing need for a robust approach to cost-effective prediction of candidate synthetic lethal interaction, particularly in cancer research. Exploiting existing public genomic databases is an ideal way to utilise existing resources with suitable sample sizes, data types, and different limitations to those of laboratory experiments. A number of computation approaches to synthetic lethality have been developed but many of these rely on data not available to cancer researchers, methods that are difficult to replicate, over-fitted to a particular dataset, having mixed validation results, or do not have a software tool accessible to the research community. These methodologies are reviewed in detail in the accompanying literature review. They will still be considered to develop an improved synthetic lethal interaction prediction tool (SLIPT).

Therefore the data types considered to be predictive of synthetic lethality and the biological questions that could be addressed by them are summarised in the Figure 2.

Figure 2. Mindmap of synthetic lethal predictors and biological areas of relevance. Underlined points have been investigated with preliminary data, italicised points are being considered in the immediate future of this project.

A bioinformatics approach has distinct limitations to experimental methods and would work well combined with genetic screen data and conventional molecular biology laboratory validation techniques to answer biological research questions. Compared with an experimental screen, a bioinformatics approach has the benefits of reduced

costs, with the potential for automation, scaling up, and replication of the same gene across populations and cell types. Analysis of public genomic data accounts for real tumour variation showing detection with tumour heterogeneity and genomic instability. Compared with a cell line or xenograft experimental model we are limited by difficulties in establishing validity of a novel method, lack of mechanism, or potential for testing drug activity in the same system. However, computational methods may further miss useful therapeutic candidates from variable genetic background and be limited by the population sampled. This research builds on previous work in an Honours project and similar approaches in the literature (Jerby-Arnon et al. 2014; Kelly 2013; Lu et al. 2015).

6.3 Simulations and Modelling Synthetic Lethality in Expression Data

Synthetic lethality was modelled for effects on expression levels and whether these are detectable in known interacting and non-interacting genes in simulated data. These were conducted for expression data but the nature of these simulations would be relevant to how synthetic lethality would manifest in other factors, particularly DNA copy number variation and DNA methylation. These simulations were discussed at length in the previous meeting and showed that synthetic lethality was detectable with our approach in simple cases. While it was less effective, the methods were able to detect synthetic lethal genes in expression data with correlation structure (generated with the multi-variate normal distribution) and were distinguishable from correlated genes. Therefore the strongest (most significant) synthetic lethal genes are more likely to be true synthetic lethal partners and a high number of hits are expected from correlated genes and co-regulated pathways.

The power of the method to detect interactions depleted with increasing multiple tests, interactions, and cryptic (third party) interacting partners. Increased sample size counteracted these effects as expected. This led the idea that pathways would be more suitable as the focus of this project. Biological pathways led to fewer multiple tests, more relevant to understanding cancer biology, and are often drug targets in practice.

6.4 Developing a Synthetic Lethal detection methodology

6.4.1 Testing Multivariate Normal Simulation of Synthetic lethality

We have developed a model of synthetic lethality in gene expression data based on sampling a Multivariate Normal distribution. This enables simulation of statistically testing for synthetic lethal genes where the true and false positives are known, discovery of the expected test statistic distributions for different conditions, educated thresholds for public data analysis, and building a complex model with known correlation structure between genes. Sampling a small number of genes from this model shows, in Figure 4, that synthetic lethality is detectable with in a simple model.

Figure 4. Chi-Square (upper) and p-values (lower) distributions show that synthetic lethal partners (red) are distinguishable from correlated (blue) and other genes (black) in an example simulation of sampling 1000 samples and 100 genes, from a multivariate normal distribution with 1 (left), 2 (centre), and 3 (right) synthetic lethal partners respectively, showing that synthetic lethal genes become more difficult to detect if there are more true partners.

Figure 5. Chi-Square (upper), FDR adjusted p-values (centre), and Holm adjusted p-values (lower) show that show that synthetic lethal partners (red) are distinguishable from correlated (blue) and other genes (black) are distinguishable replicated across 1000 replicate simulated sampling of 1000 samples and 100 genes, from a multivariate normal distribution with 1 (left), 2 (centre), and 3 (right) synthetic lethal partners respectively, showing synthetic lethal genes become more difficult to detect in with more true partners but adjusting p-values may be too stringent an approach to this.

Having shown that the Chi-Square test is capable of detecting synthetic lethality, Figure 5 shows that detecting synthetic lethality in a simple case is largely robust and reproducible across many replicates with synthetic lethal and correlated genes clearly having higher test statistic scores and lower adjusted p-values than the null distribution of non-synthetic lethal genes when there are only 1 or 2 synthetic lethal partners. While it is promising that correlated genes and synthetic lethal partners could be distinguished from other genes in a simple case, there is also indication that true synthetic lethal partners (candidates as robust drug targets) and their correlated genes (or pathways) could be distinguished by test statistic.

However, such clear evidence of synthetic lethality by co-loss under-representation is rarely detected in public data analyses, indicating cryptic additional synthetic lethal genes compensating for the loss of both the query and putative synthetic partner. Therefore higher-order synthetic lethal is potentially very common, difficult to detect, and confounding attempts to identify synthetic lethal pairs from gene expression data. In Figure 5, more than 3 synthetic lethal partners will be difficult to identify directly with a Chi-Square test. Although deeper understanding of the system could still enable use of the procedure to prioritise small numbers of candidate genes, estimate the number of underlying true synthetic partners, and identify the biological pathways interacting with a gene to focus complementary experimental approaches.

With higher number of true synthetic lethal genes there is no clear threshold for Chi-Square values (or associated p-values) to detect synthetic lethality and choosing any threshold is a trade-off between sensitivity (ensuring all true positives are detected) and specificity (reducing the number of false positives detected). Receiver operating characteristic (ROC) curves, as shown in Figures 6 and 7, summarise this trade-off to show the statistical performance of a test where the true synthetic lethal genes are known in the simulated data. Performance of a statistical test is measured as the area under the ROC (AUROC) curves, as shown in Figures 8 and 9, to compare performance across simulations for different parameters such as type of model, correlation structure, the total number of genes, sample size and number of true synthetic lethal genes. A random predictor has an AUROC of 0.5, whereas an ideal predictor has an AUROC of 1.0, so intermediate values are expected.

6.4.2 Receiver Operating Characteristic Curves

Figure 6. ROC curves showing statistical performance (by area under the curve) of a synthetic lethal simulation based on sampling a Binomial distribution, with 20,000 genes, averaged over 1000 replicates, sample size (1000, 2000, 5000, or 10,000) and number of synthetic lethal genes (up to 100) varies by panel and colour showing better performance with fewer synthetic lethal genes or higher sample size.

Figure 7. ROC curves of a synthetic lethal simulation based on sampling a Multivariate Normal distribution, with 20,000 genes, averaged over 1000 replicates, sample size and number of synthetic lethal genes varies by panel and colour showing better performance than a Binomial model and similar performance with correlation structure (upper panes).

Figure 8. Comparison of Binomial (red) and Multivariate Normal models with

(blue) and without (green) correlation structure by simulation with 1000 samples, 20,000 genes, sample size varied by pane, and number of synthetic lethal partners on the x axis where performance on the y axis is measured as the AUROC showing better performance in the Multivariate Normal model than the Binomial model and similar performance in the Multivariate Normal model with correlation structure added for all simulation parameters. There was better performance with fewer synthetic lethal partners or higher sample size with both Multivariate Normal models.

Figure 6 shows performance of an earlier model based on the Binomial distribution for gene function calls, based on similar a Normally distributed model of gene expression which called gene function from an arbitrary expression cut-off. This model is shown for comparison with Multivariate model we have chosen to develop since the Multivariate model, as shown in Figure 7, has better performance, allows the inclusion of correlation structure expected in gene expression data for biological pathways, and could have variable gene function cut-offs. The Binomial model defines the synthetic lethal condition in a way that, while ensuring at least on synthetic lethal partner is active in query deficient samples, disrupts the number of samples with functional synthetic lethal genes compared to other genes affecting the expected proportions in the Chi-Square test.

Figures 7 and 8, show that the Multivariate model which corrects this effect by specifying synthetic lethal genes differently performs better in simulations, even with correlation structure expected to disrupt the synthetic lethal detection. There is indication in Figure 8 that correlation structure even improves the performance of simulations. Although replicated across parameters, the difference in performance of simulations with correlated genes (with each synthetic lethal partner) is marginal and the number of correlated genes is still vastly outnumbered by the total number of genes (20,000 modelling a complete mammalian genome). Simulations with fewer total genes may show the impact of correlated genes more clearly, which is biologically plausible since some co-regulated pathways do involve a substantial proportion of the genome.

As indicated, the models behave as expected when performing better when simulated with higher sample size and fewer true synthetic lethal genes. As summarised in Figure 9, this behaviour occurs in simulation with all of the models discussed above. The number of synthetic lethal partners impacts performance with a sigmoidal decay where higher sample size (albeit approaching the limit of feasible genomic-scale projects) markedly delay decay of AUROC towards random 0.5. Therefore a large sample size is crucial for bioinformatics synthetic lethal discovery. Only a small number

of synthetic lethal partners will be detectable with a gene-centric approach motivating pathway-centric approaches and accounting for pathway structure, which has shown be more reproducible between model organism experiments (Dixon et al. 2009). However, whether potential false positives are more likely to be correlated genes or occur due to the sheer number of true negatives (and multiple tests) is unclear. The impact of correlation structure on the simulated data is explored in detail below in Figures 10-12 and the results of these simulations repeated is shown in Figure 13. Figure 9. Summary of effect of sample size and number of synthetic lethal partners on performance of simulations for prediction of synthetic lethality by AUROC on continuous scale (left) and as a barplot (right) showing that sample size (by colour) and number of synthetic lethal partners (x axis) affects performance as expected in which was replicated across all 3 models discussed above.

6.4.3 Simulated Expression Heatmaps

In Figures 10-12 below, simulations are summarised with expected (Sigma) and generated (Correlation) structure of gene expression patterns in the top figures. The following line shows how the expression and gene function calls have been distributed with correlation structure and ordering samples (columns) to ensure a synthetic lethal partner or query gene is active in each sample.

Figure 10. Simulation for 1 SL partner (100 genes, 1000 samples)

Figure 11. Simulation for 2 SL partners (100 genes, 1000 samples)

Figure 12. Simulation for 3 SL partners (100 genes, 1000 samples)

As shown in the Figures 10-12, the correlation structure of the simulated gene expression data (upper right) largely reflects the expected sigma matrix (upper left) used to specify the variation in the Multivariate Normal distribution with some variation due to low sampling error. The Sigma and correlation matrices show blocks of correlated genes with each synthetic lethal partner where there are 1, 2, or 3 synthetic lethal partners in Figures 10, 11, and 12 respectively. In the gene expression heatmap (lower right) and associated discrete gene function calls based on a threshold of the 30% quantile (lower left), the sample (column) ordering shows how samples were ordered so at least one synthetic lethal gene is active in all query deficient samples. The row (gene) ordering is based on a Chi-Square test statistic value and odds-ratio sign (with negative genes at the top), apart from Query gene at the top (with positive odds-ratio). The Chi-Square values are shown on the outer colour bar on a log scale and the inner colour bar annotates the known gene class in the simulation: query (blue), synthetic

lethal (red), correlated (orange), and other (green).

With 1 synthetic lethal partner, in Figure 10, the relationship between synthetic lethal (and correlated genes with the Query gene is clear and detectable with Chi-Square test (as shown with the colour bars) as expected. The relationship is clearer in the true synthetic lethal partner showing that it should be distinguishable from confounding correlated genes. With multiple synthetic lethal genes, as shown in Figures 11 and 12, the true synthetic lethal partner is less related to the expression profile of the Query gene and the co-loss under-representation is more difficult to detect since the number of co-occuring loss of synthetic lethal genes expected (even in Query functional samples is low). In these examples, the Chi-Square test still correctly identifies synthetic lethal genes with the highest test statistic, although with a less well defined cut-off and it may not be reproducible (as discussed above). This is consistent with more synthetic lethal partners being able to recover function and ensure cell survival which is plausible given the evolutionary robustness of molecular networks, difficulty detecting individual gene pairs in gene expression data, and rates of recurrence or drug resistance in cancer patients. Therefore we have to consider cryptic synthetic lethal genes compensating for Query and candidate synthetic lethal partners due to higher-order genetic redundancy, cancer genomic evolution and cellular heterogeneity.

6.4.4 Replication Simulation Heatmap

The declining performance in ROC curves with more synthetic lethal genes shows that the ability to robustly distinguish synthetic lethal genes from other genes (including their correlated genes) declines as the synthetic lethal genes do not consistently have a higher Chi-Square test statistic across replicate sampling simulations. Although it is noted that increased sample size can compensate for this decline, increasing the number of expected co-loss events and sensitivity of the procedure. The effect of total gene number, impact of correlation structure, and reproducibility of Chi-Square tests across replicate sampling simulations is explored below.

Figure 13 is composed of columns of side colour bars ordered by Chi-Square and odds-ratio sign (with Query in the corrected position at the bottom) as shown in Figures 10-12 with separate columns for repeated sampling with different parameters. Figure 13 is an example of this visualisation of simulations for a small number of genes (100) and replicates (10 each for 1 to 10 synthetic lethal partners). Even in this small simulation, we see many of the processes discussed above summarised: the effect of number of synthetic lethal genes on Chi-Square tests, power to detect synthetic lethal

and other correlated genes, decaying reproducibility and variation across replicates, lack of a clear threshold, and importance of directional conditions (e.g., odds-ratio sign) to distinguish synthetic lethal and co-expressed genes. This visualisation is an effective way to capture the simulation process and compare conditions which will be valuable for more complex correlation structure and comparison to public data Chi-Square distributions.

Figure 13. Comparison of simulation across various parameters for sampling a Multivariate Normal model for 100 genes and 1000 samples with correlation structure with 10 replicates (columns) for each number of synthetic lethal partners (1 to 10 in the top colour bar) with genes sorted by chi-squared value (and odds-ratio negative at the top) this shows preferential sorting of synthetic lethal partners (red) and correlated (orange) genes near the top (on the left) for lower numbers of synthetic lethal partners which becomes less clear or consistent across replicates for higher numbers of synthetic lethal partners, reflected in less variation in chi-square values (shown in log-scale on the right) and lack of a clear prediction threshold, however positive odds-ratio genes show no preference except for the query gene associated itself as expected.

This framework may also be useful to compare different analyses of public data and infer the true number of synthetic lethal partners from the distribution of test statistic scores. With an effective visualisation, we can further explore more complex correlation structures (as shown in the supplementary Figures S1 and S2). This will be important to develop simulated data as similar to empirical data as possible, to test whether synthetic lethal and correlated genes are robustly detectable, and discover effective drug targets (which are repeatable across a cohort, tissues or species). The impact of high-order synthetic lethality, genetic background and variation between replicates indicates that more care has to be taken interpreting experimental model systems and genomics analysis will be valuable to ensure candidate drug targets are suitable for clinical application. We show below that this visualisation scales up and shows similar effects for number of synthetic lethal genes in more replicates (Figure 14), more total genes (Figure 15), and both (Figure 16).

Figure 14. Comparison of simulation across various parameters for sampling a Multivariate Normal model for 100 genes and 1000 samples with correlation structure with 100 replicates (columns) for each number of synthetic lethal partners (1 to 10 in the top colour bar) with genes sorted by chi-squared value (and odds-ratio negative at the top) this shows preferential sorting of synthetic lethal partners (red) and correlated (orange) genes near the top (on the left) for lower numbers of synthetic lethal partners

which becomes less clear or consistent across replicates for higher numbers of synthetic lethal partners, reflected in less variation in chi-square values (shown in log-scale on the right) and lack of a clear prediction threshold, however positive odds-ratio genes show no preference except for the query gene associated itself as expected.

Figure 15. Comparison of simulation across various parameters for sampling a Multivariate Normal model for 1000 genes and 1000 samples with correlation structure with 10 replicates (columns) for each number of synthetic lethal partners (1 to 10 in the top colour bar) with genes sorted by chi-squared value (and odds-ratio negative at the top) this shows preferential sorting of synthetic lethal partners (red) and correlated (orange) genes near the top (on the left) for lower numbers of synthetic lethal partners which becomes less clear or consistent across replicates for higher numbers of synthetic lethal partners, reflected in less variation in chi-square values (shown in log-scale on the right) and lack of a clear prediction threshold, however positive odds-ratio genes show no preference except for the query gene associated itself as expected.

Figure 16. Comparison of simulation across various parameters for sampling a Multivariate Normal model for 1000 genes and 1000 samples with correlation structure with 100 replicates (columns) for each number of synthetic lethal partners (1 to 10 in the top colour bar) with genes sorted by chi-squared value (and odds-ratio negative at the top) this shows preferential sorting of synthetic lethal partners (red) and correlated (orange) genes near the top (on the left) for lower numbers of synthetic lethal partners which becomes less clear or consistent across replicates for higher numbers of synthetic lethal partners, reflected in less variation in chi-square values (shown in log-scale on the right) and lack of a clear prediction threshold, however positive odds-ratio genes show no preference except for the query gene associated itself as expected.

6.5 Simulation of synthetic lethality in graph structures

6.5.1 Developing a multivariate normal expression from graph structures

6.5.2 Simulations over simple graph structures

6.5.2.1 Performance

6.5.2.2 Synthetic lethality across graph stuctures

6.5.2.3 Performance with inhibition links

6.5.2.4 Performance with 20,000 genes

6.5.3 Simulations over pathway-based graphs

6.5.4 Comparing methods

6.5.4.1 SLIPT and Chi-Squared

6.5.4.1.1 Correlated query genes

6.5.4.2 Correlation

6.5.4.3 Bimodality with BiSEp

6.5.4.4 Linear models

6.5.5 Developing a linear model predictor of synthetic lethality

6.5.5.1 Linear models

6.5.5.2 Polynomial models

6.5.5.3 Conditioning

6.5.5.4 SLIPTv2

6.6 Significance

Development of an effective synthetic lethal discovery tool for bioinformatics analysis has a wide range of applications in genetics research including functional genomics,

medical and agricultural applications. Of particular interest is a complementary approach to discovery of synthetic lethal drug targets for cancer therapy to aid the cancer research community which currently relies on cell line and mouse models for screening and validation experiments (Fece de la Cruz et al. 2015). The potential for synthetic lethal drug design against cancer mutations including gene loss or overexpression could lead to a revolution in cancer therapy and chemoprevention with personalised treatment of cancers and high risk individuals. Examples of the synthetic lethal strategy to cancer treatment have been shown to be clinically effective with many large-scale RNAi screens underway to discover more cancer gene function and drug targets for similar application.

However, there are limitations to both experimental screens and computational approaches, both known to be prone to false-positives. Modelling and simulation of synthetic lethal discovery in genomic data has been explored to address these concerns and ensure the validity of candidate synthetic lethal interactions, particularly given the recent emergence of a number of conflicting synthetic lethal screening and prediction approaches. Exploring synthetic lethality in simulated data will ensure the optimal performance of our prediction method with comparison to the distribution of test statistic distribution in empirical gene expression data, informed selection of thresholds for prediction, and estimated error rates. The model of gene expression with known synthetic lethal genes is limited by the assumption that it represents the distribution of gene expression when it may not. Having shown synthetic lethality is detectable in simple models and added correlation structure, the model still needs to be developed to better represent real data. However, the behaviour of synthetic lethal genes and effects of parameters explored so far remains important to inform future model design and interpretation of empirical data analysis. The synthetic lethal discovery strategy could be adapted to any form of gene inactivation or disruption such as such as changes to gene expression, regulation, epigenetics, DNA sequence, or copy number which could plausibly induce cell death due to SL interactions. Further applications of synthetic lethal interactions such as analysis of gene networks, tissue specificity, evolutionary conservation, or drug target feasibility are possible with synthetic lethal candidates predicted with confidence on a large scale.

Network analysis enables properties of the network and its connectivity to be measured and compared across datasets (Barabsi & Oltvai 2004). Tissue specificity is an important consideration, largely unexplored with synthetic lethal studies, since it has clinical importance to ensure targeted drug treatments are effective, predict adverse

effects in other tissues, determine whether targeted treatments could be repurposed for other cancer types or diseases, and whether drug resistance mechanisms could emerge. Comparison of tissues, populations, and species can all ensure that synthetic lethal predictions are robust, that experimental candidates are clinically relevant, and treatments designed to exploit them would be specific to the disease in large patient cohort (with known biomarkers).

Drug targets must be feasible to have effective anti-cancer interventions designed against them, which raises the need for targets with existing drugs in the clinic, trials, or feasible to development with structural analysis or screening. Druggable targets could be selected by gene functions known to be amendable to drugs, with a structure amenable with development, with conserved specific sites without homology to other genes, or with known approval or developing drugs which could be repurposed from other disease applications.

6.7 Future Directions

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

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

research community (manuscript and code release in preparation).

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

6.8 Conclusion

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

Chapter 7

Discussion

Chapter 8

Conclusion

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`. Accessed: 22/03/2017.

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

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 brca2deficient tumours with inhibitors of poly*adpribose* polymerase. *Nature*, **434**(7035): 913–7.

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

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

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

Cancer Research UK (2017) Family history and cancer genes. `http://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/family-history-and-inherited-cancer-genes`. Accessed: 22/03/2017.

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

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

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

Chen, 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.

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

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

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.

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.

Green, R.E., Briggs, A.W., Krause, J., Prufer, K., Burbano, H.A., Siebauer, M., Lachmann, M., and Pääbo, S. (2009) The Neandertal genome and ancient DNA authenticity. *EMBO J*, **28**(17): 2494–2502.

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 grid^{TM} (caBIG^{TM}). *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.

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

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

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

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.

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.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J.K., *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**(5802): 1113–1118.

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).

Oxford Nanopore Technologies (Nanopore) (2017) Oxford Nanopore Technologies. `https://nanoporetech.com/`. Accessed: 27/03/2017.

PacBio (PacBio) (2017) Pacific Biosciences. `http://www.pacb.com/`. Accessed: 27/03/2017.

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.

Quantum Biosystems Inc. (Quantum Biosystems) (2017) Quantum Biosystems, . `http://www.quantumbiosystems.com/`. Accessed: 27/03/2017.

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.

Rothberg, J.M. and Leamon, J.H. (2008) The development and impact of 454 sequencing. *Nat Biotechnol*, **26**(10): 1117–1124.

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

The Cancer Society of New Zealand (Cancer Society of NZ) (2017) What is cancer? `https://otago-southland.cancernz.org.nz/en/cancer-information/other-links/what-is-cancer-3/`. Accessed: 22/03/2017.

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

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

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

The Internation Cancer Genome Consortium (ICGC) (2017) ICGC Data Portal. `https://dcc.icgc.org/`. Accessed: 26/03/2017.

Themo Fisher Scientific (ThermoFisher) (2017a) Ion Proton System for Next Generation Sequencing. `https://www.thermofisher.com`. Accessed: 27/03/2017.

Themo Fisher Scientific (ThermoFisher) (2017b) SOLiD Next Generation Sequencing. `https://www.thermofisher.com`. Accessed: 27/03/2017.

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

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

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

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

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

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

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

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

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

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

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

Tutt, A., Robson, M., Garber, J.E., Domchek, S.M., Audeh, M.W., Weitzel, J.N., Friedlander, M., Arun, B., Loman, N., Schmutzler, R.K., *et al.* (2010) Oral poly(adp-ribose) polymerase inhibitor olaparib in patients with brca1 or brca2 mutations and advanced breast cancer: a proof-of-concept trial. *Lancet*, **376**(9737): 235–44.

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

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

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

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

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

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001) The sequence of the human genome. *Science*, **291**(5507): 1304–1351.

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.

Wadman, M. and Watson, J. (2008) James Watson's genome sequenced at high speed. *Nature*, **452**(7189): 788.

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

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

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

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

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

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

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

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

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

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189): 872–876.

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.

Appendix A

Sample Correlation

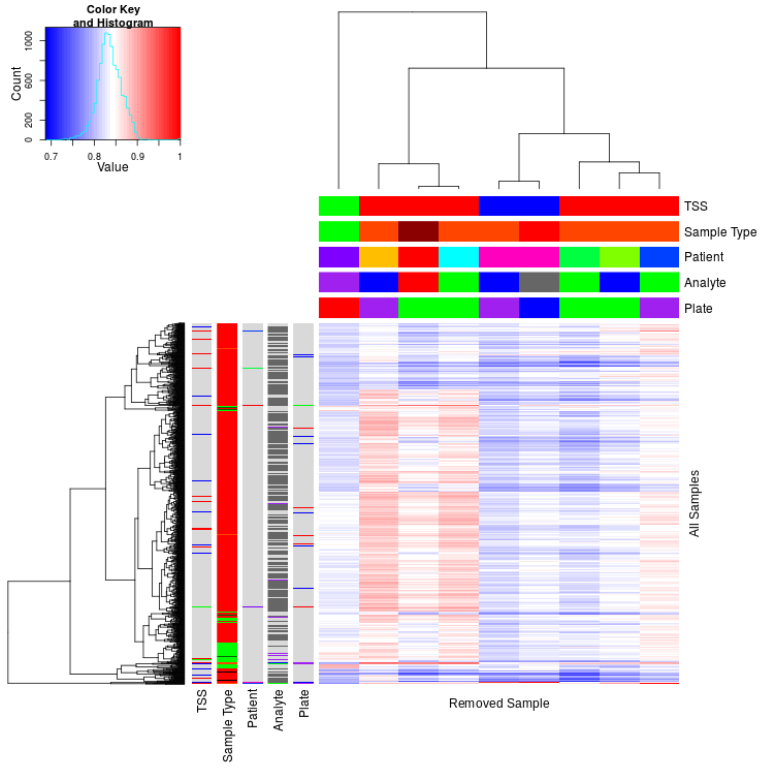


Figure A.1: **Correlation profiles of removed samples.** Correlation matrix heatmap (Euclidean distance) of all samples in TCGA breast cancer dataset (left) clustered for all samples against removed samples (top): tissue source site (TSS), sample type with reds for tumour and greens for normal, patient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

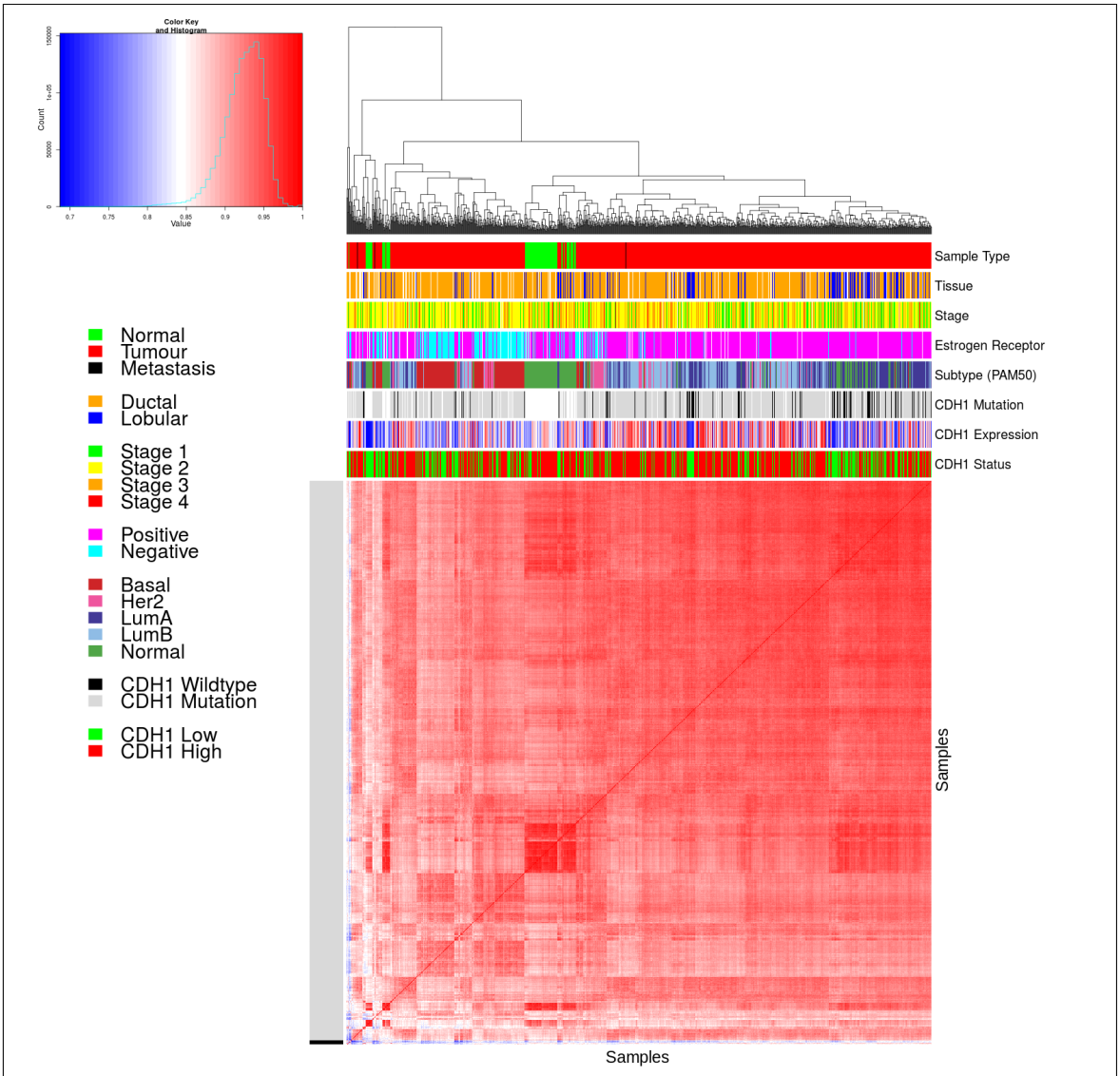


Figure A.2: **Correlation analysis and sample removal.** Correlation matrix heatmap (Euclidean distance) of all samples in TCGA breast cancer dataset against each other annotated for sample clinical data: sample type, tissue type, tumour stage, Estrogen receptor (IHC) and intrinsic subtype (from the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

Appendix B

Software Used for Thesis

Table B.1: R Packages used during Thesis

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

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

diptest	CRAN	0.75-7	0.75-7	0.75-7	
doParallel	CRAN	1.0.10	1.0.10	1.0.10	1.0.10
dplyr	CRAN	0.5.0	0.5.0	0.5.0	0.5.0
ellipse	CRAN		0.3-8	0.3-8	0.3-8
evaluate	CRAN		0.1	0.1	0.9
fdrtool	CRAN		1.2.15		
fields	CRAN		8.1		
flexmix	CRAN	2.3-13	2.3-13	2.3-13	
forcats	CRAN	0.2.0			
foreach	CRAN	1.4.3	1.4.3	1.4.3	1.4.3
foreign	base	0.8-67	0.8-67	0.8-67	0.8-66
formatR	CRAN		1.4	1.4	1.4
Formula	CRAN		1.2-1		1.2-1
fpc	CRAN	2.1-10	2.1-10	2.1-10	
futile.logger	CRAN		1.4.3	1.4.3	1.4.1
futile.options	CRAN		1.0.0	1.0.0	1.0.0
gdata	CRAN	2.17.0	2.17.0	2.17.0	2.17.0
geepack	CRAN		1.2-1		
GenomeInfoDb	Bioconductor		1.10.2	1.10.1	
GenomicAlignments	Bioconductor		1.10.0	1.10.0	
GenomicRanges	Bioconductor		1.26.2	1.26.1	
ggm	CRAN		2.3		
ggplot2	CRAN	2.1.0	2.2.1	2.2.1	2.1.0
git2r	CRAN	0.15.0	0.18.0	0.16.0	0.15.0
glasso	CRAN		1.8		
GO.db	Bioconductor		3.4.0	3.2.2	3.3.0
GOSemSim	Bioconductor		2.0.3	1.28.2	1.30.3
gplots	CRAN	3.0.1	3.0.1	3.0.1	3.0.1
graph	Bioconductor		1.52.0		
graphics	base	3.3.2	3.3.2	3.3.1	3.3.0
graphsim	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
grDevices	base	3.3.2	3.3.2	3.3.1	3.3.0

grid	base	3.3.2	3.3.2	3.3.1	3.3.0
gridBase	CRAN	0.4-7	0.4-7	0.4-7	0.4-7
gridExtra	CRAN	2.2.1	2.2.1	2.2.1	2.2.1
gridGraphics	CRAN		0.1-5		
gtable	CRAN	0.2.0	0.2.0	0.2.0	0.2.0
gtools	CRAN	3.5.0	3.5.0	3.5.0	3.5.0
haven	CRAN	1.0.0			
heatmap.2x	GitHub TomKellyGenetics	0.0.0.9000	0.0.0.9000	0.0.0.9000	0.0.0.9000
hgu133plus2.db	Bioconductor		3.2.3		
highr	CRAN		0.6	0.6	0.6
Hmisc	CRAN		4.0-2	4.0-2	3.17-4
hms	CRAN	0.2	0.3		
htmlTable	CRAN		1.8	1.9	
htmltools	CRAN	0.3.5	0.3.5	0.3.5	0.3.5
htmlwidgets	CRAN		0.8	0.8	
httpuv	CRAN	1.3.3		1.3.3	
httr	CRAN	1.2.1	1.2.1	1.2.1	1.1.0
huge	CRAN		1.2.7		
hunspell	CRAN		2.3		2
hypergraph	CRAN		1.46.0		
igraph	CRAN	1.0.1	1.0.1	1.0.1	1.0.1
igraph.extensions	GitHub TomKellyGenetics	0.1.0.9001	0.1.0.9001	0.1.0.9001	0.1.0.9001
influenceR	CRAN		0.1.0	0.1.0	
info.centrality	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
IRanges	Bioconductor		2.8.1	2.8.1	2.6.1
irlba	CRAN	2.1.1	2.1.2	2.1.2	2.0.0
iterators	CRAN	1.0.8	1.0.8	1.0.8	1.0.8
jpeg	CRAN		0.1-8		
jsonlite	CRAN	1.1	1.2	1.3	0.9.20
KEGG.db	Bioconductor		3.2.3		

kernlab	CRAN	0.9-25	0.9-25	0.9-25	
KernSmooth	base	2.23-15	2.23-15	2.23-15	2.23-15
knitr	CRAN		1.15.1	1.15.1	1.14
labeling	CRAN	0.3	0.3	0.3	0.3
lambda.r	CRAN		1.1.9	1.1.9	1.1.7
lattice	base	0.20-34	0.20-34	0.20-34	0.20-33
latticeExtra	CRAN		0.6-28		0.6-28
lava	CRAN		1.4.6		
lavaan	CRAN		0.5-22		
lazyeval	CRAN	0.2.0	0.2.0	0.2.0	0.2.0
les	CRAN		1.24.0		
lgtdl	CRAN		1.1.3		
limma	Bioconductor		3.30.7	3.30.3	
lme4	CRAN		1.1-12		1.1-12
lubridate	CRAN	1.6.0			
magrittr	CRAN	1.5	1.5	1.5	1.5
maps	CRAN		3.1.1		
markdown	CRAN		0.7.7	0.7.7	0.7.7
MASS	base	7.3-45	7.3-45	7.3-45	7.3-45
Matrix	base	1.2-7.1	1.2-7.1	1.2-8	1.2-6
matrixcalc	CRAN	1.0-3	1.0-3	1.0-3	1.0-3
mclust	CRAN	5.2	5.2.1	5.2.2	5.2
memoise	CRAN	1.0.0	1.0.0	1.0.0	1.0.0
methods	base	3.3.2	3.3.2	3.3.1	3.3.0
mgcv	base	1.8-16	1.8-16	1.8-17	1.8-12
mi	CRAN		1		
mime	CRAN	0.5	0.5	0.5	0.4
minqa	CRAN		1.2.4		1.2.4
mnormt	CRAN	1.5-5	1.5-5		1.5-4
modelr	CRAN	0.1.0			
modeltools	CRAN	0.2-21	0.2-21	0.2-21	
multtest	Bioconductor		2.30.0	2.30.0	
munsell	CRAN	0.4.3	0.4.3	0.4.3	0.4.3

mvtnorm	CRAN	1.0-5	1.0-5	1.0-6	1.0-5
network	CRAN		1.13.0		
nlme	base	3.1-128	3.1-128	3.1-131	3.1-128
nloptr	CRAN		1.0.4		1.0.4
NMF	CRAN	0.20.6	0.20.6	0.20.6	0.20.6
nnet	base	7.3-12	7.3-12	7.3-12	7.3-12
numDeriv	CRAN		2016.8-1		2014.2-1
openssl	CRAN	0.9.4	0.9.6	0.9.6	0.9.4
org.Hs.eg.db	Bioconductor		3.1.2		3.3.0
org.Sc.sgd.db	Bioconductor		3.4.0		
parallel	base	3.3.2	3.3.2	3.3.1	3.3.0
pathway.structure .permutation	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
pbivnorm	CRAN		0.6.0		
PGSEA	Bioconductor		1.48.0		
pkgmaker	CRAN	0.22	0.22	0.22	0.22
PKI	CRAN		0.1-3		
plogr	CRAN		0.1-1	0.1-1	
plot.igraph	GitHub TomKellyGenetics	0.0.0.9001	0.0.0.9001	0.0.0.9001	0.0.0.9001
plotrix	CRAN		3.6-4		
plyr	CRAN	1.8.4	1.8.4	1.8.4	1.8.3
png	CRAN		0.1-7		0.1-7
prabclus	CRAN	2.2-6	2.2-6	2.2-6	
praise	CRAN	1.0.0	1.0.0		1.0.0
pROC	CRAN		1.8	1.9.1	
prodlim	CRAN		1.5.7		
prof.tree	CRAN		0.1.0		
proftools	CRAN		0.99-2		
progress	CRAN			1.1.2	
psych	CRAN	1.6.12	1.6.12		
purrr	CRAN	0.2.2	0.2.2	0.2.2	0.2.2
qgraph	CRAN		1.4.1		

quadprog	CRAN		1.5-5	1.5-5	1.5-5
R.methodsS3	CRAN		1.7.1		1.7.1
R.oo	CRAN		1.21.0		1.20.0
R.utils	CRAN		2.5.0		
R6	CRAN	2.1.3	2.2.0	2.2.0	2.1.3
RBGL	CRAN		1.50.0		
RColorBrewer	CRAN	1.1-2	1.1-2	1.1-2	1.1-2
Rcpp	CRAN	0.12.7	0.12.9	0.12.9	0.12.7
RcppArmadillo	CRAN			0.7.700.0.0	0.6.700.6.0
RcppEigen	CRAN		0.3.2.9.0		0.3.2.8.1
RCurl	CRAN		1.95-4.8	1.95-4.8	1.95-4.8
reactome.db	Bioconductor		1.52.1	1.52.1	
reactometree	GitHub TomKellyGenetics		0.1		
readr	CRAN	1.0.0	1.0.0		
readxl	CRAN	0.1.1			
registry	CRAN	0.3	0.3	0.3	0.3
reshape2	CRAN	1.4.1	1.4.2	1.4.2	1.4.1
rgexf	CRAN		0.15.3	0.15.3	
rgl	CRAN			0.97.0	0.95.1441
Rgraphviz	CRAN		2.18.0		
rjson	CRAN		0.2.15		
RJSONIO	CRAN		1.3-0		
rmarkdown	CRAN		1.3	1.3	1
Rmpi	CRAN		0.6-6		0.6-5
rngtools	CRAN	1.2.4	1.2.4	1.2.4	1.2.4
robustbase	CRAN	0.92-7	0.92-7	0.92-7	0.92-5
ROCR	CRAN	1.0-7	1.0-7	1.0-7	1.0-7
Rook	CRAN		1.1-1	1.1-1	
roxygen2	CRAN	6.0.1	5.0.1	6.0.1	5.0.1
rpart	base	4.1-10	4.1-10	4.1-10	4.1-10
rprojroot	CRAN	1.2	1.1	1.2	
Rsamtools	Bioconductor		1.26.1	1.26.1	

rsconnect	CRAN		0.7		
RSQLite	CRAN		1.1-2	1.1-2	1.0.0
rstudioapi	CRAN	0.6	0.6	0.6	0.6
rvest	CRAN	0.3.2			
S4Vectors	Bioconductor		0.12.1	0.12.0	0.10.3
safe	Bioconductor		3.14.0	3.10.0	
scales	CRAN	0.4.0	0.4.1	0.4.1	0.4.0
selectr	CRAN	0.3-1			
sem	CRAN		3.1-8		
shiny	CRAN	0.14		1.0.0	
slipt	GitHub TomKellyGenetics	0.1.0	0.1.0	0.1.0	0.1.0
sm	CRAN	2.2-5.4	2.2-5.4		
sna	CRAN		2.4		
snow	CRAN	0.4-1	0.4-2	0.4-2	0.3-13
sourcetools	CRAN	0.1.5		0.1.5	
SparseM	CRAN		1.74		1.7
spatial	base	7.3-11	7.3-11	7.3-11	7.3-11
splines	base	3.3.2	3.3.2	3.3.1	3.3.0
statnet.common	CRAN		3.3.0		
stats	base	3.3.2	3.3.2	3.3.1	3.3.0
stats4	base	3.3.2	3.3.2	3.3.1	3.3.0
stringi	CRAN	1.1.1	1.1.2	1.1.2	1.0-1
stringr	CRAN	1.1.0	1.1.0	1.2.0	1.0.0
Summarized Experiment	Bioconductor		1.4.0	1.4.0	
survival	base	2.39-4	2.40-1	2.40-1	2.39-4
tcltk	base	3.3.2	3.3.2	3.3.1	3.3.0
testthat	CRAN	1.0.2	1.0.2		1.0.2
tibble	CRAN	1.2	1.2	1.2	1.2
tidyr	CRAN	0.6.1	0.6.1	0.6.1	
tidyverse	GitHub hadley	1.1.1			

timeline	CRAN		0.9		
tools	base	3.3.2	3.3.2	3.3.1	3.3.0
tpr	CRAN		0.3-1		
trimcluster	CRAN	0.1-2	0.1-2	0.1-2	
Unicode	CRAN	9.0.0-1	9.0.0-1	9.0.0-1	
utils	base	3.3.2	3.3.2	3.3.1	3.3.0
vioplot	CRAN		0.2		
vioplotx	GitHub TomKellyGenetics	0.0.0.9000	0.0.0.9000		
viridis	CRAN	0.3.4	0.3.4	0.3.4	
visNetwork	CRAN		1.0.3	1.0.3	
whisker	CRAN	0.3-2	0.3-2	0.3-2	0.3-2
withr	CRAN	1.0.2	1.0.2	1.0.2	1.0.2
XML	base	3.98-1.3	3.98-1.1	3.98-1.5	3.98-1.4
xml2	CRAN	1.1.1		1.1.1	1.0.0
xtable	CRAN	1.8-2	1.8-2	1.8-2	1.8-2
XVector	Bioconductor		0.14.0	0.14.0	
yaml	CRAN		2.1.14	2.1.14	2.1.13
zlibbioc	CRAN		1.20.0	1.20.0	
zoo	CRAN	1.7-13	1.7-14		1.7-13

Appendix C

Secondary Screen Data

A series of experimental genome-wide siRNA screens have been performed on synthetic lethal partners of *CDH1* (Telford *et al.*, 2015). The strongest candidates from a primary screen were subject to a further secondary screen for validation by independent replication with 4 gene knockdowns with different targeting siRNA. As shown in Table C.1, there is significant ($p = 7.49 \times 10^{-3}$ by Fisher's exact test) association between SLIPT candidates and stronger validations of siRNA candidates. Since there were more SLIPT− genes among those not validated and more SLIPT+ genes among those validated with several siRNAs, this supports the use of SLIPT as a synthetic lethal discovery procedure which may augment such screening experiments.

Table C.1: Candidate Synthetic Lethal Genes against Secondary siRNA Screen

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