Contents

List of Figures

List of Tables

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, this technology and the findings from it have considerable potential for significant impacts in the clinic and wider applications of genetics either directly or by enabling more focused genetics research on candidates selected from genomics or bioinformatics analysis. The completion of the draft Human Genome (**?**) 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 expanded upon the human genome, to populations (**?**), to cancers (Zhang *et al.*, 2011; **?**), and to deeper functional understanding (**??**). 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. Accordingly, genetics has become widely acknowledged as having an important role in cancer risk (in addition to environmental factors). Cancers arise from dysregulated 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 (**?**), with an estimated annual incidence of 14.1 million cases and annual mortality of 8.2 million people (**?**). 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 (**?**), among the highest incidence (age-standardised per capita) of cancer in the world (**?**).

While the genetic contribution to cancer risk and many of the molecular changes occurring in cancers are widely acknowledged (**???**), the majority of these findings have yet to impact on clinical practice. Diagnostics are traditionally based on pathological examination of tissue samples where histological staining for cell type, biomolecules and biomarkers continue to be widely used, although genetic and biochemical markers are being adopted for some cancer types. In general, the current standard of care includes 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 can be effective against certain cancers, particularly in early stage cancer or in patients with particular subtypes (such as acute myeloid leukaemia) which respond well to modern treatment regimens. Thus early diagnosis is important to patient survival and quality of life. National screening programs (which prioritise patients with a high risk of cancer) therefore aim to diagnose cancers early and subtypes more accurately. Therefore identification of patients with genetic variants or family histories (for inclusion in national databases) for high risk of particular cancers is an important health issue, particularly where effective interventions exist if these cancers are diagnosed early. Thus high risk individuals being regularly monitored for some cancers are sometimes offered preventative surgery or treatment for pre-cancerous tissue (**??**).

Chemotherapy is a last-resort treatment for many 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 often has severe adverse effects, a narrow therapeutic window, and is not suitable for chemopreventative application in many cases (**?**). Since surveillence preventative surgery (removing the tissue at risk of cancer) 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. These alternatives include 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 The Genetics and Molecular Biology of Cancers

Cancers involve dysregulation 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 (Guilford *et al.*, 1998; **?**; **?**). Cancer is widely viewed to be a genetic disease due to these familial cancer syndromes, hereditary risk factors, and the molecular changes occurring in cancers, including numerous cancer genes which have been identified **??**. 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 (**?**).

Hanahan and Weinberg (2000) identified several key molecular and cellular traits shared across most cancers as a rational approach to the complex changes that occur in cancer initiation and progression due to common molecular machinery underlying all cells. A cancer cell must possess 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 energy metabolism in order to survive (Hanahan and Weinberg, 2000; **?**). 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 **?**, with "intrinsic subtypes" identified characterised by estrogen receptor, *HER2*, and basal, epithelial signalling. 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 (**?**), with their relevance to prognosis (including predicting survival and response to neoadjuvant chemotherapy) demonstrated and a 50-gene subtype predictor from microarray and quantitative PCR (qPCR) analysis has been provided (Parker *et al.*, 2009; **?**). This has been further updated with the "claudin-low" subtype (**?**) 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 1.1.5.1.1) expand on the importance of molecular differences between cancers by examining molecular profiles across cancer tissue types (**?**).

The molecular variability of cancer has also been approached rationally at a pathway level with patients subgroups activating different molecular pathways reflecting differences in disease mechanisms (**?**). A robust approach to measuring pathway activation in cancer is with a"metagene" which gives a consistent signal as a consensus of expression across genes even if they are inversely correlated (Huang *et al.*, 2003; **?**; **?**). These are derived from the first principal component or eigenvector of a singular matrix decomposition, capturing the the most consistent variation across genes in a pathway or gene signature. **?** used gene signatures for 18 cellular pathways in breast cancer to define subtypes with distinct molecular pathway activity. In a meta-analysis of Affymetrix microarray expression data for 1,143 samples and 50 cell lines, unsupervised hierarchical clustering robustly defined subtypes with common homogeneous pathway activity despite variation in the specific mutations giving rise to them. These subtypes with shared pathway activity have similar molecular characteristics (such as DNA copy number), clinical properties and prognosis, building upon the intrinsic subypes (Parker *et al.*, 2009) and providing a functional interpretation for molecular stratification (**?**). The pathway-based subtypes often correspond to intrinsic subtypes (Gatza *et al.*, 2014; **?**) and provide finer molecular stratification such as environmental stress response (to hypoxia within HER2-enriched cancers) (Gatza *et al.*, 2011) and have pathway-specfic DNA copy number variation or essential genes (Gatza *et al.*, 2014). Gatza *et al.* (2014) extend these analyses include 52 pathway signatures from previous publications in breast cancer, replicating known characteristics (such as hormone re) of each subtype and identifying novel subtype-specific driver genes of proliferation by analysis of microarray expression from 476 from The Cancer Genome Atlas (TCGA). In addition to distinct biological functions driving growth of breast cancer subtypes, these molecular subtypes provide a rational approach based

on molecular properties to cancer treatment with combination and targeted therapies (Gatza *et al.*, 2014; Hanahan and Weinberg, 2000; **?**).

Cancer is a major health concern with a well-established genetic contribution, in risk and in the molecular changes occurring during progression (**?**). 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 **?** 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 (**?**) has transformed genetics research including the study of health and disease (**??**). 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 (**?**). These projects serve as an excellent resource for genetics research globally, particularly for cancers where genomics investigation have been widely applied to different tissues across molecular profiles Zhang *et al.* (2011); **?**); **?** . Genome sequencing technologies continue to improve, decrease in price, and become increasingly feasible in a wider range of research and 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 (**??**). 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 vertebrates) was much lower than most estimates for a genome of it's size and the number of repetitious 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 (**??**).

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, **?** conceded that genomic information alone was not sufficient for biological understanding and that many investigations re-

mained to carried out, 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.

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 created widespread anticipation for potential applications in healthcare, agriculture, ecology, conservation, and evolutionary biology despite few of these being realised yet.

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 often involve genetic mutation or dysregulated 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 (**?**).

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 polymorphism (RFLP) or repetitious microsatellite regions. These exploited sequence variation at target sites of restriction enzymes or measured the length of repetitious regions, using polymerase chain reaction (PCR), restriction enzymes, and gel electrophoresis to measure deoxyribonucleic acid (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 (**?**) enabled DNA sequencing and genotyping at a widespread scale, being less technically difficult than the Maxam-Gilbert sequencing by degradation method (**??**), 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 (**?**). 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 qPCR is another adaptation of genetic technologies to quantitatively study nucleic acids, often reverse transcribed "deoxyribonucleic acid (cDNA)" or messenger "messenger ribonucleic acid (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 and are still used for smaller scale experiments and as a "gold standard" for measuring gene expression (**?**). This also represents a shift in the application of qPCR 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, ribonucleic acid (RNA) expression, and DNA methylation levels. Due to results of comparable or arguably better quality from these newer technologies (**????**), this "gold standard" status has begun 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 a 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 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 nucleic 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 of 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 there is a low computational cost of mapping to a reference compared to 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 regarding large data and compute requirements to handle the raw output data. Compared with 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 developments in sequencing technologies. One of the main computational challenges is the assembly of 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 for 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 where 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 to target the mRNA by binding to the poly-A tail (RNA-Seq) or 5' cap (CAGE-Seq). However, this has the potential to exclude miRNA (miRNA) and long non-coding ribonucleic acid (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 outweigh 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 technical 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, insertions or deletions, 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 with 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, miRNA, lncRNA, or chromatin immunopreciptation sequencing (ChIP-Seq). RNA-Seq of the transcriptome is a common adaptation where RNA is reverse transcribed and sequenced from the resulting cDNA. This is utilised to 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 nucleic acid may be extracted for sequencing such as the coding regions of DNA (for the "exome"), the mRNA 5′ cap (CAGE-Seq), mRNA 3′ poly-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 applied to proteins 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 Sequencing Technologies

Illumina sequencing (developed by Solexa and later acquired by Illumina) was 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 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. Their Hi-Seq platforms were used in large-scale gemomics projects (such as the cancer genome atlas discussed in Section 1.1.5.1) to generate the sequence data used throughout this thesis.

NGS technologies continue to be refined with Illumina (and competitors) continuing to release refined productions, offer additional genomics-based services, and decreases overhead and operating costs to enable a wider range of research projects. As such RNA-Seq for examining the transcriptome of an organisms and for expression studies in diseases is a growing field of research and expression data will continue to be generated for a range of samples in many healthy and diseased tissues. The technology be also be further improved developments in speed and accuracy (such as Ion Torrent plat-

forms) and towards long reads of single molecules (such as Pacific Biosciences, Oxford Nanopore, and Quantum Biosystems Japan).

Due to such benefits of sequencing over previous technologies (and their continued refinement), this thesis has focused on gene expression 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 resources.

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 (**??**) and to human populations with an interest variability between individuals and health or disease risk (**??**).

Genomics databases have also focused on faciltating distribution of genomic data generated by researchers, rather than generating it themselves. GenBank hosted by the National Center for Biotechnology Information (NCBI) in the US, the The European Nucleotide Archive (ENA) hosted by the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ) hosted by the National Institute of Genetics (NIG) do so by serving as public repositories of DNA sequence data. Gene Expression Omnibus (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, serving as a resource to support ongoing research where larger datasets than would were previously accessible for many individual laboratories are available (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

The importance of genomics technologies in the future of cancer research was noticed, even in the early days of genomics (**?**). 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 research 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 (**?**), 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 (**?**), a database and website of cancer genes. This launched with 66,634 samples and 10,647 mutations from initial investigations into *BRAF*, HRAS, *KRAS*, and NRAS (**?**). 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 (**?**). 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) (**?**). They first set out to demonstrate the pilot project on brain (**?**), ovarian (**?**), and squamous cell lung (**?**) 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.

1.1.5.1.1 Findings from Cancer Genomes

The Cancer Genome Atlas pilot projects (**???**) serve to demonstrate the power of applying genomics technologies to cancer research at such as scale. In addition to sequencing 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 (**?**), 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 signalling in brain cancer. An analysis of 489 serious ovarian cancers (**?**) 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*, *BRCA2*, *RB1*, and *CDK12*. Four transcriptional subtypes of ovarian cancers were identified, alterations in *BRCA1*, *BRCA2*, and *CCLE* had an impact on patient survival, and the homologous recombination, NOTCH and FOXM1 signalling pathways were involved in ovarian cancer growth. The genomics of 178 squamous cell lung cancers (**?**) 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 aberrant 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 (**?**), 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 (**?**), 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.

1.1.5.1.2 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) (Bass *et al.*, 2014; TCGA, 2012; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**). In addition to disregarding the tissue-based distinction between colon and rectal cancers based on molecular similarlity (**?**), 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 (Bass *et al.*, 2014; TCGA, 2012; **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**).

The "Pan cancer" project (**??**) 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 (**?**) 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; **?**; **?**), 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; **?**).

1.1.5.1.3 Cancer Genomic Data Resources

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 (**???**). These serve as a vast resource of common and rare cancer types and are publicly available for further analysis (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 (**?**). 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 (**?**).

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" (**?**). 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 (**?**), which can be detected in genomics data (**??**). Important "driver" cancer genes (**?**) 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 (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 occurring in each tumour usually affecting cell fate, survival, or genome maintenance (**?**).

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 treatments. Personalised medicine was described in contrast with current strategies of health advice, screening, prognostics, and treatments based on what works well for the majority of the population. Adverse effects of these treatments occur in a significant subpopulation, particularly demographics under-represented in clinical studies.

The importance of genomics is still recognised in translational cancer research. Applications are particularly emphasised in molecular diagnosis, prognosis, and treatments 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 supercede 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 the medical expertise required for treatment. 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 treatments 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 addition to (or in lieu of) 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; **?**). 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 International Cancer Genome Consortium (ICGC) projects, which combined molecular profiles across tissue types **?**. The molecular subtypes could feasibly be included in clinical 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 that target 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 (**?**).

Despite controversy over their prohibitively high cost (**?**), targeted therapeutics have been applied as monoclonal antibodies against oncogenes (such as *HER2*) with relative sucess in clinical trials (**?**), 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 have 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 synergy between inhibitors of the oncogenes $BRAF^{\mathrm{V600E}}$ and *EGFR* in colorectal cancer is an example of such a system **?**.

Despite successful application of vemurafenib against $BRAF^{\mathrm{V600E}}$ in melanomas **??**, colorectal cancers with $BRAF^{\mathrm{V600E}}$ mutations have poor prognosis and lack drug response. **?** used an RNA interference (RNAi) screen and found that *EGFR* inhibition is synergistic with vemurafenib aginst $BRAF^{\mathrm{V600E}}$ in colon cell lines and xenografts due to feedback activation of *EGFR*. Vemurafenib induced rapid reactivation of MAPK/ERK signalling via *EGFR* in colorectal cell lines in a tissue-specific manner **?** and may be be relevant to acquired resistance in melanoma **?**. Thus combination therapies against several molecular pathways may be necessary to anticipate acquired resistance **?** and targeted therapeutics may be further refined from understanding the pathway structure and functional interactions across 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 signalling, and metabolic reactions of are all inter-related and may each be perturbed by aberrations in gene function occurring in cancers. These relationships can be represented by biological networks by mapping pairs of genes with a particular relationship. Due to the complexity of a cell, these molecular networks are very large consisting of thousands of nodes comprised by genes or proteins.

The properties of large networks were first studied by constructing random networks by randomly linking a fixed number of nodes (**??**). 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 (**??**). The small world property, made popular by findings in social networks (**?**), 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. **?** 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 **?**.

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' **?**. Hubs further differentiate into 'party' hubs (which interact simultaneously with many partners) and 'date' hubs (which interact with different partners in different conditions) **?**. 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.

? 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 **?**. 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 (**??**).

? 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 area under the reciever operating characteristic (AUROC) of 0.784 **?**. 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 **?**. As an alternative interpretation, (**?**) 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 of 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 **?**. 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 developing drugs with a desired target profile designed for the target topology Barabási and Oltvai (2004); **?**. 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 **?**. 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); **?**. Further optimisation of timing and dosing of drug combinations may increase efficacy of treatments with low efficacy applied separately. Low doses and drug holidays are other counter intuitive approaches which may increase clinical efficacy, reduce adverse effects, and reduce drug resistance (**??**).

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 conceived 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 be 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 receiving 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 (**?**). 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's definition lends itself to quantitative traits and more broadly encompasses synthetic genetic interactions (synthetic genetic interactions). These have become popular for studies in yeast genetics and cancer drug design (Boone *et al.*, 2007; **?**).

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 **??**.

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

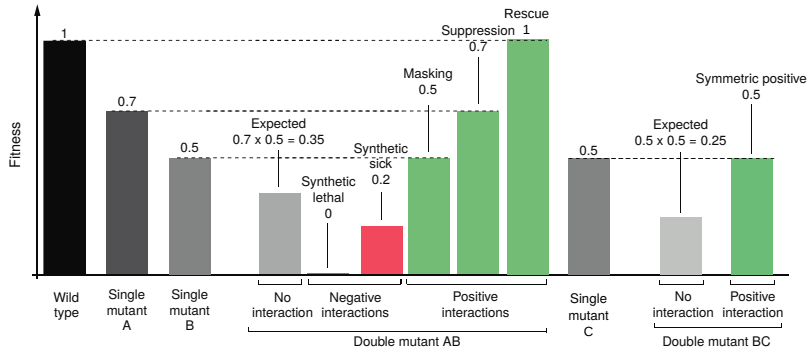

Figure 1.1: **Synthetic genetic interactions.** Impact of various negative and positive SGIs: negative interactions involve deleterious (sick) or inviable (lethal) phenotypess whereas positive interactions involve restoring viability by masking or suppressing the other mutation or complete rescue of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

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 such as chemicals, essentiality of synthetic lethal genes can be induced by the environment of a cell. An environmental stress condition 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" (**?**). 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). Many genes of the same pathway may be fucntionally 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 may occur at pathway level or 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; **?**) 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 4× more interactions detected with synthetic gene array mating screens than protein-protein interactions yeast-2-hybrid studies (**?**). 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 (**??**). 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 (**?**). 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 were relatively more inclined to have SGIs between distinct biological pathways.

1.2.3.1.1 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 (Dixon *et al.*, 2008). 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*) (**?**). 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 (**?**).

A modest 5% of interactions were conserved between unicellular (*S. cerevisiae*) and multicellular (*C. elegans*) organisms. However, the nematode SGI network had similar scale-free topology and modularity despite differences in methodology: metazoan RNAi screens are incomplete knockouts whereas screening null mutations is feasible in yeast (**?**). The nematode SGI screen identified network hubs with important interactions to orthologues of known human disease genes (**?**). 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 (**?**). 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 findings 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 (**?**). 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; **?**). Since synthetic lethality

affects cellular viability by indirect functional relationships between genes, it is suitable for indirectly targeting mutations 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 cancer cells adapt to particular oncogenic growth signals and become reliant on them to remain viable (Luo *et al.*, 2009; **?**), synthetic lethal partners of inactivated tumour suppressors are required to maintain cancer cell viability and proliferation. As such cancers are subject to "non-oncogene addiction" and are feasible anti-cancer drug targets.

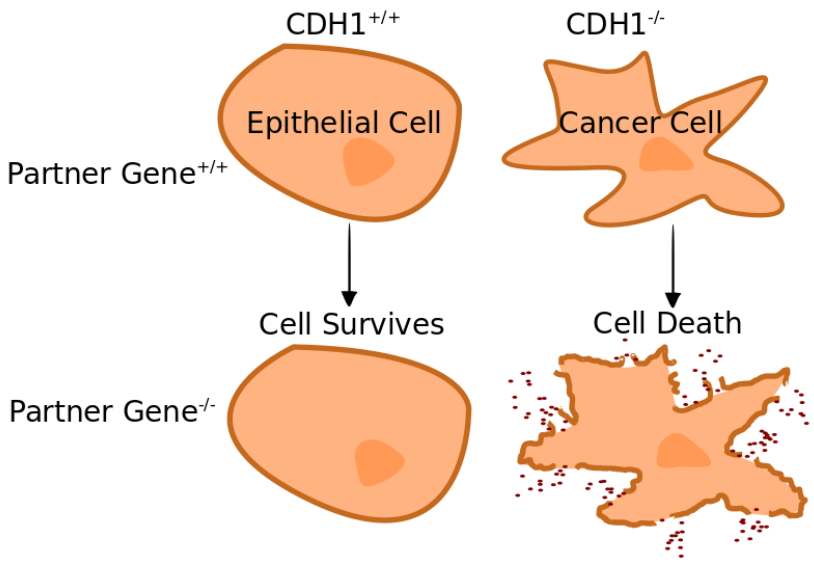

Figure 1.2: **Synthetic lethality in cancer.** Rationale of exploiting synthetic lethality for specifity against a tumour suppressor gene (e.g., *CDH1*) while other cells are spared under the inhibition of a partner gene.

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 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 (**?**). 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 **??**. *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*) (**?**). 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, $PARP1^{-/-}$ knockout mice are viable and healthy indicating low toxicity from PARP inhibition (**?**).

? 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. **?** 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 (**?**). This supports the potential for PARP inhibition as a chemo-preventative alternative to prophylactic surgery for high risk individuals with *BRCA1* or *BRCA2* mutations (**?**). 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). **?** 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 (**?**) and sporadic ovarian cancer (**?**). AstraZeneca has reported phase II trials showing the treatment is effective in *BRCA1* or *BRCA2* deficient breast (**?**) and ovarian cancers (**?**) 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. These PARP inhibitors are FDA approved for some cancers **?**, are effective against germline and sporadic *BRCA1* or *BRCA2* mutations, and are a potential prevention alternative to prophylactic surgery for high risk mutation carriers **?**.

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 (**?**). The targeted therapy is effective in both sporadic and hereditary *BRCA1* or *BRCA2* deficient tumours acting against an oncogenic molecular aberration across several tissues.

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 (**?**). The emerging RNAi technologies have vastly expanded the potential for studying genetic redundancy in mammalian experimental models including testing experimentally for synthetic lethality (**?**). Identifying synthetic lethality is crucial to study gene function, drug mechanisms, and design novel therapies (**?**). 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 (**?**). While biologically informed hypotheses have had some success in synthetic lethal discovery (**???**), interactions occurring indirectly between distinct pathways would be missed (Boone *et al.*, 2007; **?**). 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 (**??**).

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 (**?**). 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 CRISPR/Cas9/sgRNA genome editing technology (**?**), episomal gene transfer (**?**), 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 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. **?** performed an RNAi screen for synthetic lethality between *PIM1* overexpression and gene knockdown in RWPE prostate cancer cell lines. *PLK1* gene knockdown and drug inhibition was 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 in which *PIM1* and *PLK1* are co-expressed and correlated with tumour grade. Thus RNAi screening was valuable to identify 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*). **?** 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 consistent 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. **?** 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 (**?**). Therefore, *WEE1* inhibitors are an example of treatment which could be repurposed with the synthetic lethal strat-

egy. 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. **?** 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 were 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 the specificity of drugs to a particular target, and many of these drugs are not yet clinically available yet anyway, although they are clinical trials for other diseases or limited to access by patients from a particular 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 and drug screens renders meta-analysis ineffective.

Genome-scale synthetic lethal experiments (across gene pairs) are not feasible, even in model organisms, and they typically focus on specific gene candidates or the partners of a gene of interest (such as importance in health). Therefore a computational approach is more suitable for this task and may further augment experimental screening to replicate screen candidates beyond experimental models.

1.2.7 Computational Prediction of Synthetic Lethality

1.2.7.1 Bioinformatics Approaches to Genetic 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 **?** and **?** and summarised in Table **??**. 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 specificity 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 specificity of protein-protein interactions (PPIs) , an important predictor of genetic interactions, is difficult given that high-throughput two-hybrid screens occur out of cellular context for multicellular organisms.

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

1.2.7.2 Comparative Genomics

A comparative genomics approach by **?** used the results of well characterised high-throughput mutation screens in *S. cerevisiae* as candidates for synthetic lethality in humans (**?????**). 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 (**?**) were matched to human orthologues, with over 500 involving a cancer gene (**?**). 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 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 **?**. In their relational database, called "Syn-lethality", 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 function 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 developing 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" (**?**).

Machine learning approaches have also been proposed for synthetic lethal discovery (**??**). Due to concerns that these may be subject to overfitting or noise, **?** developed a meta-analysis method (based on the machine learning methods in Table **??**) for synthetic lethal gene pairs relevant to developing selective drugs against human cancer, building upon their previous database (**?**). They 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 **?**, PPI, protein complexes, and biological pathway) with strong statistical 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 (**?**) and MetaSL (**?**) 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 **?**, although interactions between pathways may be 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 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 since they lack homologues 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 (**?**). 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, **?** used modelling of signalling pathways to identify synthetic lethal interac-

tions between known drug targets and cancer genes by simulating gene knockdowns. A computational approach was 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 (**?**) for 28 signalling proteins and Gene Ontology (GO) (GO) pathways **??**. This approach successfully detected many known essential genes in the human gene essentiality database, known synthetic lethal partners in the Syn-Lethality database (**?**), 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 contained many *TP53* and AKT synthetic lethal partners, genes known to be important in many cancers, however these also have a severe impact on the signalling pathways in their essentiality 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 **?**. 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 (**?**) and Cancer Cell Line Encyclopaedia (CCLE) (Barretina *et al.*, 2012) 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 lescale-freethal 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 anticancer 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 the DAta mIning SYnthetic lethal identification pipeline (DAISY), a data-driven approach for genome-wide analysis of synthetic lethality in public cancer genomics data from TCGA and CCLE (Barretina *et al.*, 2012). 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 either essential in screens or not deleted with copy number variants. Co-expression is also used for synthetic lethality prediction based on studies in yeast (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 has not been publicly released for replication.

Although the DAISY procedure performs well and has been well received by the scientific community (**???**), 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 two 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 genome evolution of cancers, 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 **?** and **?**. 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 release 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 molecular profiles. They released this solution as the BImodal Subsetting ExPression (BiSEp) R package **?** 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 demonstrated on an analysis of 881 cell lines from CCLE (Barretina *et al.*, 2012), 442 cell lines from Catalogue Of Somatic Mutations In Cancer (COSMIC) (**?**), and RSEM normalised RNA-Seq data for 178 TCGA lung patient samples (**?**). 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 homogeneous 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 or human genomes (Lu *et al.*, 2015; Wappett *et al.*, 2016; **?**; **?**), 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 cancers. 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 signalling. 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 occurring 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 (**??????**). 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 (**???**).

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 including 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 signalling 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 signalling 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 (**?**). 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 (**???**), it is likely that EMT will drive growth of E-cadherin deficient cancers (**???**). 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 (**????**). Individuals carrying a null mutation in *CDH1* have 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, occurring 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 (**?**) 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 (**?**). 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 (**?**). 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 (**??**) detected 994 distinct mutations in 10,143 tumour samples (at a rate of 7.52%), **?** detected 632 distinct mutations in 43,865 tumour samples (at a rate of 1.71%), and detected mutations in 13.2% of 53 of the NCI60 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.

? reports *CDH1* mutations in 40 cancer tissue types including stomach (11.40% in 1342 samples), breast (10.29% in 3343 samples), large colon (2.87%), skin (2.83%), endometrial (2.81%), and bladder (1.9%) cancer. **?** reports *CDH1* mutations in 29 cancer tissue types including skin (23.41% in 598 samples), breast (14.50% in 1696 samples), ovary (13.98% in 93 samples), and stomach (11.41% in 289 samples) cancer samples. *CDH1* mutations are reported at similar rates in breast and stomach cancer in other cancer genomics projects and studies across distinct populations. **?** reports *CDH1* mutation prevalence in stomach cancer at 16.7% (**?**, 30 samples), 15% (**?**, 100 samples), 14.1% (**?**, 78 samples), and 9.4% (**?**, 393 samples). **?** also reports *CDH1* mutation prevalence in breast cancer at 12.7% (**?**, 963 samples) and 10.8% (**??**, 2051 samples). The rare plasmacytoid bladder cancer subtype also has a high prevalence of *CDH1* mutations in **?** 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-occurring 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*. **?** reports the prevalence of the mutations in these genes at 10% for *CDH1*, 49% for *TP53*, 22% for *PIK3CA* in stomach cancer (**?**, 393 samples). There is no evidence of significant co-occurring 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*. **?** also reports the prevalence of the mutations in these genes at 13% for *CDH1*, 32% for *TP53*, 36% for *PIK3CA* in breast cancer (**?**, 963 samples). There is evidence of significant co-occurring 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 short interfering ribonucleic acid (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 GO analysis of the candidate synthetic lethal partners the primary siRNA screen. This was supported by validation with Pertussis toxin, known to target $G_{\alpha i}$ signalling (**?**), as were various candidate cytoskeletal pathways by inhibition of Janus kinase (JAK) 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 available from them have immense 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 **?** 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.

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.

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

Thesis Aims

- To develop a statistical approach to detect synthetic lethal gene pairs in cancer from expression data

- To apply this methodology to public cancer gene expression data against *CDH1* and analyse pathway structure with comparisons to experimental screen data

- To construct a statistical model of synthetic lethality in multivariate normal expression data

- To develop a simulation pipeline of expression with pathway structure on a high-performance computing cluster

- To examine the statistical performance of the methodology with simulated expression including pathways and compare it to other approaches

- To release the synthetic lethal detection methodology and pathway simulation procedure as R software packages

References

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

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

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.

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

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.

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.

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.

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.

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.

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

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

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.

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.

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

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

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.

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

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.

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

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.

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

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

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

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

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

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

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

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.

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

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

Heiskanen, M., Bian, X., Swan, D., and Basu, A. (2014) caArray microarray database in the cancer biomedical informatics grid$^{\text{TM}}$ (caBIG$^{\text{TM}}$). *Cancer Research*, **67**(9 Supplement): 3712–3712.

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

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

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.

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.

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

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

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

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

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

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

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.

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.

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.

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.

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

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.

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

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

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.

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

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

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

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.

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.

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.

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

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

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

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

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

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

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., 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.

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.

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

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.

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.