Library Declaration Form

University of Otago Library

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

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

Degree: Doctor of Philosophy

Department: Department of Biochemistry

Permanent Address: 710 Cumberland Street, Dunedin, NZ

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

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

i) a library

ii) an individual

at the discretion of the University of Otago.

Signature:

Date:

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

S. Thomas Kelly

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand.

31 March 2017

Abstract

Background

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

Approach

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

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

Findings

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

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

of directionality between the candidates was detected.

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

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

Publications during Candidature

Several publications have been prepared during the candidature of this thesis, including some findings related to the thesis topic on which we elaborate in more depth here. Please see the original article for the results which have been accepted for publication by peer-review:

Kelly, S. T. and Spencer, H. G. (2017) Population-Genetics Models of Sex-Limited Genomic Imprinting. *Theoretical Population Biology*

Kelly, S. T., Chen, A., Guilford, P. J., and Black, M. A. (2017) Synthetic lethal interaction prediction of target pathways in E-cadherin deficient breast cancers. (Manuscript in Preparation for *BMC Genomics*)

Software Packages during Candidature

Several software packages in the R language have been released on GitHub while preparing this thesis. Please see the appropriate GitHub repository for more information on installing and running these packages, on the following account: `https://github.com/TomKellyGenetics`

`slipt` is the Synthetic Lethal interaction Prediction Tool, released to accompany the synthetic lethal publication above. `slipt-app` contains an application developed in the R `shiny` environment as part of a related project.

Several plotting functions were customised for the Figures in this thesis (and the above publications), notably `heatmap.2x` and `vioplotx` have been prepared largely for use during this project but are also documented and available to other R users. These are enhancements to the CRAN `gplots` and `vioplot` packages respectively and are intended be user-friendly for those familiar with `heatmap.2` or `vioplot` and `boxplot` (base R) functions. These are backwards compatible with these functions, taking similar imputs as demonstrated in the appropriate vignettes.

The use of iGraph (the R `igraph` package) operations of graph-network structure in the analysis and simulations of pathways involved several original or customised functions to manipuate or plot `igraph` objects and adjacency matrices. These can be install separately from their respective repositories of with the metapackage: `igraph.extensions`. `plot.igraph` enables plotting graph networks with customised inhibitor arrow and node colours.`info.centrality` enables the calculation of additional node and network centrality metrics not available in the `igraph` package. `pathway-.structure.permutation` enables testing of related states or node groups in a network by directionality of shortest paths. The `graphsim` package has been set up to simulate a multi-variate normal gene expression dataset with `mvtnorm` while deriving the correlation structure, Σ, from a graph structure. Note that these require various packages for graph theory, statistics and matrix operations and these will be installed as dependencies.

Acknowledgements

I thank my supervisors A/Prof. Mik Black and Prof. Parry Guilford for their support and guidance throughout this and prior projects. It has been a great experience, I hope to keep in contact following my studies at Otago and look forward to seeing what else your research groups produce in the future.

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

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

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

your time when all I really needed was a chat over a walk or a pint and a moment to think clearly.

I must also thank various organisations supported this research project:

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

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

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

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

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The thesis presents research into genetic interactions based on genomics data and bioinformatics approaches. Here we introduce the recent developments in genomics and Bioinformatics, particularly in their applications to cancer research. Synthetic lethal interactions are a long standing area of interest to genetics in both model organisms and cancer biology. A bioinformatics approach to synthetic lethal interactions enables much wider exploration of the potential inter-connected nature than previous candidate-based approaches. An alternative approach is experimental screening which will be presented and contrasted with bioinformatics approaches in more detail in the literature review chapter. We outline some of the reasons why these interactions are of interest in fundamental and translational biology but we must first define these and similar interactions. A particularly novel application of synthetic lethal interactions is design of treatments with specificity against loss of function mutations in tumour suppressor genes. We focus on E-cadherin (encoded by *CDH1* as a prime example of this and as such briefly review the role of this gene in cellular and cancer biology.

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, there are many ways in which it could impact upon clinical and wider applications of genetics either directly or by enabling more focused genetics research from candidates selected from genomics or bioinformatics analysis. The publication of the Human Genome marks a major accomplishment in genetics research and also opens a new chapter of challenges to utilise this genomic scale information effectively. Technologies in

this area have only accelerated in development since then and many global large-scale projects have attempted to expand from the human genome, to populations, to cancers, and to deeper functional understanding. However, impact on the clinic has been slower than initially anticipated following the completion of the "draft" genome with much research ongoing or yet to be done before genomics technologies become widely adopted in healthcare and oncology. Here we outline the main genomics technologies and bioinformatics approaches which have led to availability of genomics data techiques used in this thesis with particular interest in those with applications in cancer research or the clinic in the future.

1.1.1 The Human Genome Revolution

The advent of the Human Genome sequence has transformed genetics research in several ways (Lader *et al.*, 2001). Systematic, unbiased studies across all of the genes in the genome are viable in ways not considered before this time. The undertaking and success of a large, international project has set an example for numerous projects to follow, many being genomics investigations into other species or expanding to the functional or population level. These projects serve as an excellent resource for genetics research globally, particularly for cancers where it has been widely explored for different tissues across a range of molecular profiles. Genome sequencing technologies continue to improve, dropping in price and becoming more feasible in the research lab and for clinical applications.

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 vertbrates) was much lower than most estimates of a genome of it's size and the number of repetious transposon elements was very high. Even the figure of 30-40,000 genes given by the original publication is now widely viewed as an overestimate.

Accounting for the "complexity" encoded by the human genome with so few genes has led to investigations into the molecular function, expression profile, and population variation. When publishing the genome, the author's themselves concede that genomic

information alone is not biological understanding and that there is much that remains to be done, with their main goal being to share the raw genome data is available for further inquiry rather than interpreting it themselves. While genomics technologies and genomics projects have flourished since then, the need in turn for systematic means of interpreting data of such scale and for the interdisciplinary expertise to do so has only grown.

The project originally set out to isolate sections of the genome with labour intensive cloning techniques and individually sequencing DNA with a known locus in the genome by the Sanger methodology scaled up with the capillary sequencing approach. However it became apparent that this would take an incredibly long time and the project shifted to the "shotgun" sequencing approach: cutting the genome into many small sections and reconstructing their locus by aligning them by overlapping sections after the sequencing was performed. While sequencing technologies have changed since this project, this paradigm shift in sequencing at the genome-scale stands, with the "shotgun" approach being the norm for most sequencing. This represents a shift in genome biology from meticulously tracing the cloned DNA segments to relying on computational approaches to handle the data afterwards, to either assemble genomes *de novo* or map DNA segments to a known reference sequence. This approach was largely successful with the majority (80%) of the genome being sequenced over 15 months, a relatively short period in the project which began a decade before the announcement of the draft sequence (covering 94% of the genome). However, it has some shortcomings including handling the repetious regions of a genome.

While some follow-up work has been performed to improve the quality of this sequence, map the distance between contiguous sequences, and "close the gaps", repetious regions including the central (centromere) and peripheral (telomere) parts of the chromosomes remains to be mapped. This remains a challenge in modern genomics but for most purposes, the non-repetious aspects of the genome amenable to "shotgun" sequencing are sufficient to study the regions of functional importance (such as genes) and of variation between individuals. For this reason nucleotide variation has largely superceeded repetious elements in studies of genetic variability and are used far more widely than structural variants. Genomes continue to be published with incomplete assembles (and accepted to be completed) if the contigs are large enough to be useful for most intents and purposes, particularly if the unknown sequences between contiguous sequences are known. As such, physical distance (length in base-pairs) has largely superceeded genetic distances based on breeding and reference genomes are widely used

to faciltate genetics experiments and for wider genomics applications such as mapping genetic variants or expressed genes. Further resources have been developed to enable access to the human genome data such as the "genome browsers" provided online by the National Centre for Biotechnology Information (NCBI), University of California Santa Cruz (UCSC), and Emsembl jointly hosted by the European Bioinformatics Institute and the Sanger Institute.

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

Expectations of Genomics

The human genome attracted a high public profile, particularly with the idea of "junk DNA", an unexpectedly large proportion of the genome which did not appear to be functional. This DNA not encoding genes has since been found to be rich in other functional elements, including microRNA, long non-coding RNA, and regulatory elements such as sites of epigenetic modifications. Genomics has stimulated investigations into many of these previously largely explored areas of functional genetics and thus been of immense value in genetics research, attracting high expectations for further applications. Genomics research has become anticipated for it's potential for widespread applications in healthcare, agriculture, ecology, conservation, and evolutionary biology.

An area of particularly high interest for the clinical impact of genomics is oncology, with potential application across cancer diagnostics, prognosis, management, and developing treatment. Cancers are diseases characterised by uncontrolled cell growth, often driven by genetic mutation or dyregulated gene expression. However, as with

many areas of genomics, direct impact of genomics on the clinic has been limited compared to initial expectations following the publication of the human genome and compared with widespread adoption in cancer research. At the time of the genome announcement, it was expected that genomics would become widespread in the clinic and some popular science writers even humoured the idea of "home genomics", analogous to how personal computing became adopted by the public. This was largely intended to be for healthcare applications.

Despite significant advances in genomics technologies, including an unprecedented decrease in costs, the most immediate benefit of genomics to patients is indirect usage of the technologies to identify specific biomarkers and drugs to become adopted into clinical practice: diagnostic testing and pharmacological treatment. Clinical adoption of genomics directly raises far more difficult translational challenges. A key issue is ensuring comparable reliability to current genotyping, gene expression, and molecular pathology based on technologies such as polymerase chain reaction (PCR), Sanger Sequencing, or antibody staining. The debatable issue of incidental findings much also be addressed: that is, how to handle or report the additional genetic data gathered from a genome other than that intended to be tested for, such as variants with unknown, potential, or even established malignant implications for the health of the patient.

Along with the overhead costs of genomics being prohibitive to personal usage, the computational demands and genetics expertise required to assemble and interpret a genome have made genomics still largely used for research purposes by institutions. However, some companies are offering direct-to-consumer genetics testing, pushing for public awareness of genetic risks, testing of outwardly healthy people, and preventative medicine. Due to ethical and legal concerns, companies (such as 23andMe) offering genetic testing directly to consumers without clinical consultation have been restricted to reporting on traits for interest and ancestry rather than for healthcare.

1.1.2 Technologies to Enable Genetics Research

DNA Sequencing and Genotyping Technologies

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

is laborious and limited to well cahracterised variable regions of the genome, generally genes or nearby marker regions.

The Sanger (dideoxy) chain termination method enabled DNA sequencing and genotyping at a wider scale than previously possible. This quickly became more widely adopted over the Maxam-Gillert sequencing by degradation method develop around the same time due to the technique being less technically difficult and requiring less radioactive and toxic reactants, which have since been replaced by flourescent dyes. Another advantage of the Sanger methodology is the 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 as a "gold standard" to validate the findings of newer approaches.

Microarrays and Quantitative Technologies

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

Microarrays represent a truly high-throughput molecular technique, reducing the cost, time, and labour required to study molecular factors such as genotype, expression,

or methylation across many genes, making it feasible to do so over a statistically meaningful number of samples. Microarrays are manufactured with probes which measure binding of particular nucleotide sequences to either quantitatively detect the presence of a sequence such as a single nucleotide polymorphism (SNP) or quantify for DNA copy number, heterozygousity, expression (cDNA), or methylation (bisulfite treated DNA) purposes. Microarray technologies have popularised "genome scale" studies of genetic variation and expression.

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

A recently developed alternative to these approaches for gene expression is the "nanoString" technology which also samples a selected panel of genes. However, it lacks the scale of microarrays, being effectively limited to 800 probes. This technology differs by giving an absolute measure of the number of transcripts rather than relative measures between genes within a sample, with the manufacturers claiming a higher accuracy. While promising for studies of known genes, such as biological pathways and cancer genes, this technology has not been widely adopted due to higher cost to alternatives and higher-throughput technologies without bias to known genes being feasible. A similar refinement to the qPCR approach, "droplet" or "digital" PCR also offers to produce absolute measures of transcript abundance. These technologies may become more widely adopted for candidate gene panel research studies and clinical testing with higher data quality as the cost becomes less of an issue. However, the higher throughout of microarray technologies enables a more unbiased approach to test a large number of genes for differential expression for example.

Massively Parallel "Next Generation" Sequencing Technologies

Similar to microarrays, the introduction massively parallel sequencing technologies have further expanded the availability of high-throughput molecular studies to researchers, with corresponding availability of genomics data from these studies. This

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

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

However, there are also additional challenges arising from using data generated from such a recent innovation. This includes ethical issues such as the ongoing debate on how to handle the "incidental findings" which may arise from sequencing on such vast scale, particularly with regard to whether NGS technologies are suitable for clinical use and "variants of unknown significance", those with undetermined or contested health implications. The methodology itself also has some 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 RNA experiments in Eukaryotes where it is commonplace target the mRNA by binding to

the poly-A tail (RNA-Seq) or 5' cap (CAGE-Seq). However, this has the potential to exclude microRNAs (miRNA) and long non-coding RNAs (lncRNA) of interest unless the sample is prepared specifically to study these. Similarly capturing a subsection of the genome for an "exome" or reduced representation bisulfite sequencing (RRBS), focuses on sequencing DNA sequences and methylation levels of CpG sites near known genes to reduce cost, noise, and incidental findings.

In many cases, the benefits of NGS technologies over microarrays still outwiegh the additional cost. NGS is highly adaptable to different applications: DNA sequencing (whole genome or exome), DNA methylation (bisulfite-Seq), RNA-Seq, miRNAs, lncRNA, or chromatin immunopreciptation (CHIP-Seq). NGS scales to all genes and beyond for these molecular applications without having to design new probes as required for a microarray. Thus NGS technologies are not limited to genes already characterised sequence or functions, do not need to be updated with new probes for each genome annotation release, and do not require a reference genome at all for new species. A "transcriptome" can be assembled *de novo* for an expression study in any organism.

NGS technologies also have the advantage of greater potential accuracy and sensitivity than microarrays, depending on the sequencing depth or "coverage", theoretically sensitive down to the exact number of molecules for each transcript. NGS experiments are regarded as "reproducible" with no need for technicals replicates, although these are still performed for a subset of samples in many projects for quality assurance purposes. NGS has a wider dynamic range than microarrays: able to detect SNPS, indels (frameshifts), and splice variants in addition to quantifying DNA copy number or transcript abundance.

Applying NGS technologies varies in cost depending on the platform but is generally substantially more the costly than a microarray experiment for gene expression, limiting the number of sample sizes feasible in many studies. However, many NGS platforms now support barcoding to label samples and "multiplex" their sequencing to perform several samples at once to reduce time and cost of reagents with a sacrifice of read depth. In many cases, this approach is sufficient to compare the expression across many samples and bioinformatics methods are able to correct for varied read depth between samples. Furthermore, refinements of NGS sequencing technologies, the economies of scale, and emerging sequencing technologies have the potential to further reduce the cost of sequencing to the point where it may become feasible for widespread clincal application.

There is ongoing technology in development to overcome the various drawbacks to established NGS technologies. These emerging technologies, sometimes called "3 generation" sequencing aim to introduce radically different approaches to sequencing with distinct advantages. Long reads are the focus of several technologies, accuracy and read length of NGS platforms has improved over time but it is still difficult to assemble or map highly repetious sequences. Another refinement is the sequencing of single molecules in real time, with the potential benefits of low input material and studying 3D structure of nucleic acids. Many of these technologies focus on improving the quality and accuracy of sequences, with higher throughput, read depth, more accurate methodology, avoiding PCR bias or sequencing RNA directly for quantitative studies. Another benefit to highly sensitive sequencing platforms is the potential application in forensic, ancient DNA, and single cell samples where the amount or quality of nucelic acids is low. Single cell applications are of particular interest in cancer research due to the heterogeneity of cells within tumours and their role in diagnostics or drug resistance.

Established Sequencing Technologies 454 sequencing (acquired by Roche) commercially released from 2005 to 2013 was the first NGS technology, generating a vast 1 million reads per day or 400-600Mbp in a 10 hour run. This technology used the "pyrosequencing" method of sequencing by synthesis, detecting phosphates released when a compatible nucleotide reacts and extends the DNA synthesis of a complementary strand. This technology popularised NGS with the first complete genome from a single individual (James Watson, 2007) and the Neanderthal ancient DNA studies (Svaante Paabo, 2006 & 2009). While this technology was capable of reads up to 1kb, reads of 400-500bp were more typical and the error rate was higher for sequences of the same base consecutively. This is still relatively long reads for an NGS technology but it has been discontinued due to competing short read technologies being more cost-effective with lower running costs.

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

Illumina sequencing (developed by Solexa and later acquired by Illumina) was also

released in 2006. It utilises reversible terminating dyes to sequence by synthesis with a lower accuracy (98%) and read lengths of 150-250bp. Illumina more than makes up for relatively short reads (along with improving the read length of the technology) and low accuracy with high-throughput and cost effectiveness, with a Hi-Seq 2500 platform producing up to 3 billion reads (600Gbp) in a 3-10 days run. Illumina has further reduced the cost of sequencing with the economies of scale with the HiSeq 10X claiming to produce a human genome 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.

Emerging Sequencing Technologies Ion Torrent (also acquired by Life Technologies) released in 2010 employs "sequencing by synthesis" but in a drastically different way with ion semiconductor sequencing, detecting H^+ ions released when bases during DNA synthesis. Without the use of optical detection, the Ion Torrent system is compact offering rapid, cost-effective sequencing with the potential to scale with the future development of silicon semiconductors with have historically doubled in density every 2 years (Moore's Law). It is capable of reads of 100-200bp in only an hour (as fast as 4 seconds per base) and up to 400bp in a 2 hour run with an accuracy of 99.6% (dropping to 98% for consecutive sequences of 5 bases). While fast, cost effective, and accurate, Ion Torrent has short reads and modest throughput (depending on the platform 100Mbp to 32Gbp) compared to other sequencing technologies.

Pacific Biosciences (PacBio) released the RS and RS II platforms in 2010 and 2011 (now acquired by Roche) to make up for the short reads in NGS technologies with the single molecule real time (SMRT) approach capable of long read lengths, averaging between 2.5-7kb and up to 80kb. The PacBio methology traps each molecule in a zero mode waveguide (ZMW) and sequences it in real time. The RS II has 150,000 ZMW and an output of 500Mbp-1Gbp per SMRT cell (doubling that of the RS), with the capacity to run up to 16 concurrently for 0.5-6 hours. While the single molecule sequencing approach has strengths in sensitivity and potential to detect 3D stuctures, such as G-quadruplexes, this has the drawback of slowing down the sequencing and reducing the throughput of the platform. Another issue is sequence quality with the raw data as poor as 20-30%. However, PacBio recommends specific software to assemble as consensus with 99.999% (Q_{50} for sequences with over 20x coverage, regardless of sequence repeats

or GC composition. Despite concerns over data quality and higher cost than other approaches, the long reads are appealing for genome assembly and in many genome studies combine PacBio reads with more accurate short read technologies. However, due to the poor separate quality of reads this technology may not be appropriate for RNA-Seq studies, while it does have the potential for high sensitivity and detecting alternative splicing were it be improved. PacBio has recently released the Sequel (2016) system, increasing the throughput of the SMRT Cells 7x to 1 million ZMW holes.

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

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

continue to improve, becoming more feasible for a wider variety of applications.

Due to such benefits of sequencing over previous technologies (and their continued refinement), this thesis has focused on RNA-Seq data in contrast to prior studies on microarray data. 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.

Bioinformatics as an Interdisciplinary Approach to Genomics Data

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 to genetics to cell modelling, neuroscience, cancer development, ecology, and evolution.

In practice, many researchers identify with both fields 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 and Bioinformatics reseources.

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

1.1.3 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. The International HapMap Project, 1000 Genomes Project, and the 100K Genome Project aim to gather genetic variation data across human populations, along with gathering clinical and environmental variables for health and disease association studies. Whereas the ENCODE, ModENCODE, and FANTOM projects aim to characterise the functional aspects of human and model organism genomics. ENCODE in particular has attracted much criticism for over-inflated claims of DNA functionality. A notable finding of the ENCODE projects is that a high number of DNA sites bind to proteins or are transcribed into RNAs. However, these are not necessarily of functional importance. Conversely, the FANTOM projects approach this problem by focusing on expressed mRNAs, microRNAs, and epigenetic marks in each tissue or cell type. An area of recent interest are the long non-coding RNAs, the focus of FANTOM6, the next phase of the project amenable to the CAGE-Seq technologies developed in prior FANTOM projects.

Other genomics databases have focused on faciltating distribution of genomic data generated by researchers, rather than generating it themselves. Genbank (NCBI) in the US, EMBL in Europe, and the DDBJ (NIG) in Japan do so by serving as repositories of DNA sequence data. GEO, arrayExpress, and caArray serve a similar purpose as a resource for gene expression datasets. These serve as a resource to support ongoing research to ulitise data for genes of interest to particular research groups and further to make inferences based on larger datasets than accessible to any individual laboratory.

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. Reverse transcribed RNAs are a common such adaptation, employing RNA-Seq to the transcriptome. This is ulitised to be quantify the levels of RNA and identify which regions of DNA are expressed. Similar bisulfite treatment converts cytosine residues to uracil (sequenced as thymidine), sparing methylated cytosine enabling it to be distinguished with bisulfite-Seq for high-throughput detection of the notable epigenetic mark and generating an epigenome. High-throughput gel and mass spectrometry techniques have been employed to proteins and metabolites to generate the proteome and metabolome respectively. In this way, so called "omics" profiles across a wide range of biomolecules in a cell are produced in many experimental laboratories.

SUch genomics technologies have since been applied to single cell isolates and to detect traces of foetal or tumour molecules in blood or urine.

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.

1.1.4 Cancer Genomes

It's importance in the future of cancer research was noticed, even in the early days of genomics (Dickson, 1999). The Cancer Genome Project (CGP) based at Wellcome Trust Sanger Institute in the UK were among the first to launch investigations into cancer after the publication of the Human Genome, using this genome sequence, consensus across the cancer reserach literature, and sequencing the genes of cancers themselves. Initially, the Sanger Institute set out to sequence 20 genes across 378 samples while the Human Genome project was still ongoing (Collins and Baker, 2007), optimising sequencing and computation infrastructure for a larger project while doing so. The main aim of the Cancer Genome Project was to discover "cancer genes", those frequently mutated in cancers by comparing the genes of cancer and normal tissue samples, both "oncogenes" and "tumour suppressors" which are activated and inactivated respectively in cancers. This project is ongoing and the UK continues to be involved in international sequencing initiatives and those focused on particular tissue types.

The Sanger Institute also hosts the Catalogue of Somatic Mutations in Cancer (COSMIC), a database and website of cancer genes. This launch with 66,634 samples and 10,647 mutations from initial investigations into BRAF, HRAS, KRAS2, and NRAS (Bamford, 2004). It has since expanded to include 1,257,487 samples with 4,175,8787 gene mutations curated from 23,870 publications, including 29,112 whole genomes (Release v79 (23/08/2016 `http://cancer.sanger.ac.uk/cosmic`). This database now also identifies cancer genes from DNA copy number, differential gene expression and differential DNA methylation.

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 (2008), ovarian (2011), and lung (2012) cancers. In 2009, the project expanded aiming to analyse 500 samples each for 20-25 tumour tissue types. They have since exceeded that goal, with data available for 33 cancer types including 10 "rare" cancers, a total of over 10,000 samples.

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

The International Cancer Genome Consortium

TCGA and the Cancer Genome project in the UK are part of a larger International Cancer Genome Consortium (ICGC), now a concerted effort across 16 countries to sequence the genome, transcriptome, and epigenome of 50 tumour types from over 25,000 samples total. With some redundancy the following countries are profiling various tumour types: USA (including TCGA), China (16), France (10), Australia (4), South Korea (4), the UK (4), Germany (4), Canada (3), Japan (3), Mexico (3 in collaboration with the US), Singapore (2), Brazil, India, Italy, Saudi Arabia, and Spain. This is inherently international and several projects are collaborations, such as between the USA and Mexico, Australia and Canada, Singapore and Japan, along with the UK and France representing the European Union. In order to avoid competing the existing TCGA projects, some countries focus on a particular cancer they have health interest: Australia (melanoma), Brazil (melanoma), India (oral), Saudi Arabia (thyroid), and Spain (CML). Others focus on a particular tissue subtype with poor prognosis: The UK (triple negative or Her2+ breast cancer), France (clear cell kidney), Australia and Canada (ductal Pancreas). Another approach is to focus on rare or child cancers: Canada, Italy, France, Germany, Japan and Singapore, and the US (TARGET project). Particularly countries in Asia (China, Japan, Singapore, and South Korea883 samples)

have emphasised the value of adding tumour data from non-Western countries or non-European populations in addition the data from Europe and the TCGA in the US. Data from 9 of these countries is already available on the ICGC website with the project ongoing.

Subpopulation and Cell Line projects

Similarly, the San Antonio Cancer 1000 Genome Project also focuses on the genomics and clinical data of their local population. Another more specific project is the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) a joint project between Canada and the UK focusing particularly on breast cancer. The METABRIC project gathered the genome and transcriptome of 2,433 samples to identify driver mutations and breast cancer subtypes. Other projects have extended the genomics technologies to profile cancer cell lines used in research, including the Cancer Cell Line Encyclopaedia (CCLE) and the COSMIC Cell Lines projects. The CCLE run by the Broad Institute and Novartis has analysed 883 cell line samples including DNA copy number, array expression, mutaions, and drug response data. The COSMIC project aims to profile 1000 cell lines by exome, copy number, expression (RNASeq), DNA methylation.

These projects represent a massive ongoing global effort to understand cancers at the molecular level across the genome and beyond. They serve as a fantastic resource for further analysis, being publicly released, regularly updated, and making sample sizes far higher than feasible in prior microarray studies. The potential of genomics in cancer research is widely recognised as is the value of technological and bioinformatics advances to enable it to continue to do so.

1.1.5 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.

Cancer Genes and Driver Mutations

Personalised or Precision Cancer Medicine

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

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

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

Molecular Diagnostics and Pan-Cancer Medicine

Gene Expression Signatures and Biomarkers

Targeted Therapeutics and Pharmacogenomics

Targeting Oncogenic Driver Mutations

Tissue Specificity and Genetics Background Effects

Network Biology, Network Medicine, and Polypharmacology While targeted therapeutics have shown some levels of success for drug discovery, particularly in anticancer applications, exploiting the connectivity of molecular networks or designing combination therapies are alternatives that may yield more effective results using a network pharmacology framework (Hopkins 2008). Rational design of drugs selective to a single target has largely failed to deliver clinical efficacy. In fact, many existing effective drugs modulate multiple proteins and were selected for biological effects or clinical outcome rather than molecular targets. Network biology and polypharmacology (specific binding to multiple targets) are a possible way to develop drugs with a desired target profile designed for the target topology. Multi-target treatments aim to achieve a clinical outcome through modulation of molecular networks since the genetic robustness of a cell often compensates for loss of a single molecular target.

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

Thus a network understanding of a cell 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 PARP inhibitors or $BRAF^{V600E}$ and EGFR inhibitor combination therapy, could enable the understanding of the mechanisms of such interventions which exploit genetic interactions, protein interactions, cell signalling or gene regulation 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 a vast potential to improve cancer medicine, particularly expanding the application of targeted therapeutics against molecular targets to inactivation of tumour suppressors and other genes that are difficult to target directly. Synthetic lethal interactions have been studied for their implications to gene function and drug mode-of-action in model organisms for a long time. Here we introduce the concept of synthetic lethality as it was originally concieved and how it has been adopted conceptually in cancer research. Detecting this interactions at scale and interpreting them is the focus of much of this thesis, here we start with an overview of the concepts involved, initial work on the interaction, and the rationale for applications to cancer. Meanwhile specific investigations in to synthetic lethality in cancer and detection by experimental screening or computational analysis will be covered in the literature review chapter.

1.2.1 Synthetic Lethal Genetic Interactions

Genetic interactions are a core concept of molecular biology, discovered in some of the earliest experiments investigating Mendelian genetics. Epistasis was biologically defined as the effect of an allele at one locus may mask the phenotype of another (Bateson & Mendel 1909). Statistically, epistasis is defined by significant disparity between the observed and expected phenotype of a double mutant, compared to the respective phenotypes of single mutants and the wild-type (Fisher 1919). Fisher's broader definition lends itself well to quantitative traits and also encompasses Synthetic genetic interactions (SGIs) which have become popular for studies in yeast genetics and cancer drug design (Boone *et al.* 2007; Kaelin 2005).

Synthetic genetic interactions are substantial deviations from the expected null mutant phenotype (usually measured by organism or cellular growth or viability) rather than the expected additive effects of the single mutants. Unlike biological epistasis, this deviation does not necessarily constitute a single mutant phenotype (as shown in Figure 1). SGIs of interest tend to be more viable than either single mutant or less viable than the expected double mutant. In negative SGIs, mutations are 'synergistic', resulting in more deviation from the wild-type than expected. For yeast viability phenotypes, the terms 'synthetic sick' (SSL) and 'synthetic lethal' (SL) refer to negative SGIs giving growth inhibition and inviability respectively. In cancer research synthetic lethality is used to describe any negative synthetic genetic interaction with specific killing or

growth inhibition of a mutant cell, even though these may be formally classed as SSL interactions. In positive SGIs, mutations are 'alleviating', resulting in less deviation from the wild-type than expected. For yeast viability phenotypes, the terms 'synthetic suppression' (SS) and 'synthetic rescue' (SR) are used to refer to positive SGIs giving partial or complete restoration of the wild-type growth rate from single mutants with growth impairment and lethal phenotypes respectively. Yeast studies have found that negative SGIs are markedly more common than positive SGIs (Tong *et al.* 2004); this has since been replicated in a number of model systems (Boucher & Jenna 2013).

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

Synthetic Lethal Drug Design

1.2.2 Synthetic Lethal Concepts in Genetics

Functional Genetics

Conditional and Induced Essentiality

Functional Redundancy

Evolutionary and Developmental Biology

Genetic Robustness and Network "Re-wiring"

Cancer and Translational Biology

Non-Oncogene Addiction

1.2.3 Synthetic Lethal Concepts in Genetics

Synthetic Lethal Pathways

Experimental Inference

Models and Computational Detection

1.2.4 The Potential of Synthetic Lethality for Anti-Cancer Medicine

Rationale of Exploiting Synthetic Lethality in Cancers

Indirect Targeting and Tumour Suppressor Genes

Synthetic Interactions and Gene Dosage

Homology and Indirect Targeting for Anti-Cancer Specificity

1.2.5 Prior Studies on Synthetic Lethality

Experimental Studies of Synthetic Lethality

Genetic and chemical high-throughput screens have already 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 cancer syndromes. 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 novel computational methods show the need for synthetic lethal discovery in the fundamental genetics and translational cancer research community. However, existing computational methods are complex, error-prone, and difficult to understand, interpret or adopt for biologically trained

researchers. 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.

A high-throughput method, the synthetic gene array (SGA), has been developed specifically to test for SGIs in the budding yeast *Saccharomyces cerevisiae* (Tong *et al.* 2001). Automated mating experiments are used to generate and measure the viability of haploid double null mutants for a query gene and a library of ˜4700 deletion mutants. A proof of concept experiment with 8 query genes found SGIs within and between biological pathways (Tong *et al.* 2001). This approach was later extended to 132 query genes to experimentally map ˜4100 SGIs between ˜1000 genes with each query gene involved in an average of 34 SGIs (0.6% of the genome) with high variation: up to 146 partners (Tong *et al.* 2004). They noted that this was 4-fold higher than interactions discovered with yeast-2-hybrid studies of protein-protein interactions. The network is scale-free with power-law vertex degree distribution and low average shortest path length (3.3).

These observations are consistent with a larger network (˜170,000 SGIs out of 5.4 million gene pairs) constructed from SGA analysis of 1712 query genes in *S. cerevisiae* which showed high functional organisation, genetic redundancy, and pleiotropy (Baryshnikova *et al.* 2010b; Costanzo *et al.* 2010). Hub genes were 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. The modularity of the *S. cerevisiae* SGI network is pronounced enough to predict function of novel genes (Costanzo *et al.* 2011). The supplementary data for this network has been updated to support SGA data for 1897 query genes.

Boone *et al.* (2007) predicted ˜200,000 SGIs in the yeast genome of ˜6200 genes; the number of genes individually essential for life is vastly outnumbered by the number of genes essential in combination due to synthetic lethality. This is consistent with the chemical genomic screens finding 97% of yeast genes are conditionally essential for optimal growth in some environmental stress condition (Hillenmeyer 2008). Together this gives an evolutionary rationale for the abundance of SGIs and surprisingly low number of essential genes in a genome due to the genetic redundancy and network robustness of a cell.

The SGA methodology has been adapted to screen deletion or gene knockdown libraries for SGIs using SpSGA with mating in fission yeast, *Schizosaccharomyces pombe (Dixon et al. 2008)*; using eSGA or GIANT-coli with conjugation in bacteria, *Es-*

cherichia coli (Babu *et al.* 2014; Butland *et al.* 2008; Typas *et al.* 2008); or using RNA interference (RNAi) in the nematode worm, *Caenorhabdiits elegans* (Lehner *et al.* 2006; Tischler *et al.* 2008). The SGA methodology has also been modified for the analysis of essential genes in *S. cerevisiae* by using hypomorphic promoter replacement alleles rather than (lethal) null mutants (Davierwala *et al.* 2005). The SGA method has been combined with microarray technologies for the high-throughput diploid based synthetic lethality analysis of microarrays (dSLAM) and epistatic miniarray profiles (EMAP) approaches which enable pooled mating experiments and increased sensitivity in gene subsets respectively in *S. cerevisiae* (Collins *et al.* 2006; Ooi *et al.* 2003; Schuldiner *et al.* 2005; Schuldiner *et al.* 2006). The EMAP approach has also been successfully adapted for comparative analysis of the distantly related *S. pombe* (Roguev *et al.* 2008; Roguev *et al.* 2007). Davierwala *et al.* (2005) used conditional (tet regulated) or temperature sensitive alleles to study essential genes with SGA, whereas Schuldiner *et al.* (2005) developed the decreased abundance of mRNA perturbation (DAmP) method for high-throughput generation of partial disruption in essential genes. This enables SGI screening in various species at the genomic-scale, including essential genes, by gene knockdown or non-lethal mutation.

Davierwala *et al.* (2005) found that the SGI network (30 query genes against 575 essential genes) for essential genes was 5x denser than the non-essential network in *S. cerevisiae (Costanzo et al. 2010; Tong et al. 2004).* Therefore essential genes are SGI network hubs and are involved in the majority of SGIs, despite comprising only ˜18% of the *S. cerevisiae* genome. Essential pathways are therefore highly buffered consistent with strong selection for survival with partial loss of essential gene function. The essential gene network also presented with scale-free topology and rarely contained interactions found in a protein-protein interaction network. While interacting essential genes were likely to have related functions, this was less evident than in non-essential networks. Around the same proportion of essential genes are found in other Eukaryotes indicating that mammalian SGI networks may have similar levels of SGIs.

While distantly related, by millions of years of evolution, the budding yeast *S. cerevisiae* and the fission yeast *S. pombe* SGI networks are comparable (Dixon *et al.* 2009b). These yeasts are both single-celled, were analysed by similar SGA methodology with colony growth phenotypes, and used null or hypomorphic mutation to inactivate genes. The RNAi based SGI screens in *C. elegans* or *Drosophila melanogaster* are less comparable to the yeast models due to differences from phenotyping diploid multicellular organisms and the potential residual wild-type gene activity in an RNAi experiment.

Relatively poor conservation of specific SGIs in *S. cerevisiae* were replicated in *S. pombe* (222 queries against a library of 2663 genes); however, around ˜29% of the interactions tested in both species form a conserved SGI network (Dixon *et al.* 2008). The rest of the interactions reveal species specific differences expected in distantly related species, however many of the species specific interactions were still conserved between biological pathways, protein complexes, or protein-protein interaction modules. Negative SGIs were likely to be conserved between biological pathways, whereas positive SGIs were more likely to be conserved within a pathway or protein complex (Roguev *et al.* 2008). Conservation of pathway redundancy was also found between *S. cerevisiae* and prokaryotes with *E. coli* experiments (Butland *et al.* 2008).

RNA interference in Eukaryotes Despite reasons difficulties comparing yeast SGA data and metazoan RNAi screens for SGIs, ˜5% of interactions in *C. elegans* were conserved in *S. cerevisiae,* and the nematode SGI network showed similar scale-free topology and modularity (Bussey *et al.* 2006). The nematode SGI screen also identified network hubs and implicated their importance via interaction with orthologues of known human disease genes (Lehner *et al.* 2006). However, genetic redundancy at the gene or pathway level fails to account for the lack of direct conservation of SGIs between yeasts and nematode worms, consistent with an induced essentiality model of SGIs which allows for conservation of gene function with network restructuring over evolutionary time (Tischler *et al.* 2008). While Eukaryotic 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.

Examples of Clinical Impact

Examples of gene interactions include the synthetic lethal interaction of BRCA1 or BRCA2 with PARP1 in breast cancer and the unexpected synergistic effect of $BRAF^{V600E}$ and EGFR inhibitors in colon cancer. PARP inhibitors are one of the first targeted therapeutics against a tumour suppressor mutation with success in clinical trials. $BRAF^{V600E}$ and EGFR inhibitors block a feedback regulation loop and enable use of BRAF inhibitors beyond melanomas. Thus we can develop targeted anticancer therapeutics which exploit complex interactions to distinguish normal and cancerous cells.

Investigation of such interactions at a network level could enable deeper understanding of their impact on cancers and prediction of novel drug targets.

Many genes are lost in cancer and yet few interventions target these tumour suppressor mutations compared to targeted therapies for gain of function mutation in oncogenes. Synthetic lethality, also known as 'non-oncogene addiction' or 'induced essentiality' in the context of cancer therapy, is a powerful design strategy for therapies selective against loss of gene function with potential for application against a range of genes and diseases. There are several examples of clinically relevant applications of genetic interactions including specific targeting of mutations in BRCA tumour suppressor genes with PARP inhibitors by inducing synthetic lethality in breast cancer (Farmer *et al.* 2005) and the synergy between inhibitors of the oncogenes $BRAF^{V600E}$ and EGFR in colorectal cancers (Prahallad *et al.* 2012).

BRCA and PARP genes demonstrate the application of the synthetic lethal approach to cancer therapy (Ashworth 2008; Kaelin 2005). BRCA1 and BRCA1 are homologous DNA repair genes which have been popularised for their role as tumour suppressors; mutation carriers have high risk of breast and ovarian cancers. The BRCA 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 (Bryant *et al.* 2005).

Bryant *et al.* (2005) showed that $BRCA2^{-/-}$ cells were sensitive to PARP inhibition by siRNA of PARP1 or drug inhibition (which targets PARP1 and PARP2) using Chinese hamster ovary cells, MCF7 and MDA-MB-231 breast cell lines. This effect was sufficient to kill mouse tumour xenografts and showed high specificity to BRCA2 deficient cells in culture and xenografts. Farmer *et al.* (2005) replicated these results in embryonic stem cells and showed that $BRCA1^{-/-}$ cells were also sensitive to PARP inhibition relative to the wild-type with siRNA and drug experiments in cell culture and drug activity against BRCA 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 BRCA 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 to clinical uses against BRCA mutations in

hereditary and sporadic cancers (Ashworth 2008; Kaelin 2005). PARP inhibition has been found to be effective in cancer patients carrying BRCA mutations and some non-BRCA mutant ovarian cancers, suggesting synthetic lethality between PARP and other DNA repair pathways (Ström & Helleday 2012). This supports the potential for PARP inhibition as a chemo-preventative alternative to prophylactic surgery for high risk individuals with BRCA mutations (Ström & Helleday 2012). Hormone-based therapy has also been suggested as a chemo-preventative in such high risk individuals and aromatase inhibitors have completed phase I clinical trials for this purpose (Bozovic-Spasojevic *et al.* 2012). Ström and Helleday (2012) also postulate increased efficacy of PARP inhibitors in the hypoxic DNA-damaging tumour micro-environment.

A PARP inhibitor, olaparib, showed fewer adverse effects than cytotoxic chemotherapy and anti-tumour activity in phase I trials against BRCA deficient familial breast, ovarian, and prostate cancers (Fong *et al.* 2009) and sporadic ovarian cancer (Fong *et al.* 2010). AstraZeneca has reported phase II trials showing the treatment is effective in BRCA deficient breast (Tutt *et al.* 2010) and ovarian cancers (Audeh *et al.* 2010) with a favourable therapeutic window and similar toxicity between carriers of BRCA 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 BRCA genes in the tumour; unsurprisingly, these patients, benefit most from the PARP inhibitor treatment and have increased platinum sensitivity in combination treatment. This demonstrates the clinical impact of a well characterised system of synthetic lethality with known cancer risk genes. Synthetic lethality has the benefit of being effective against inactivation of tumour suppressor genes by any means, broader than targeting a particular oncogenic mutation (Kaelin 2005). The targeted therapy is effective in both sporadic and hereditary BRCA deficient tumours acting against an oncogenic molecular aberration across several tissues.

Oncogene targeted therapies have also been developed but have problems with resistance, recurrence, tissue specificity, and design of a drug which selectively inhibits the oncogenic variant rather than the normally functional proto-oncogene. BRAF is a serine/threonine kinase gene in the MAP kinase pathway with several oncogenic mutations including the V600E mutation implicated as a driver in some colorectal, hairy cell leukaemia, lung, melanoma, non-Hodgkin's lymphoma, and thyroid cancers (Entrez Gene). EGFR is the epidermal growth factor receptor gene which detects extracellular signal ligands including EGF and TGFα. Mutations resulting in overexpression

of EGFR are oncogenic in brain, colorectal, and lung cancers (Entrez Gene). Both oncogenes have been explored in some detail as separate drug targets with successful development of drugs acting against $BRAF^{V600E}$ in melanoma and drugs against EGFR in lungs cancers, however they have limited efficacy separately in colorectal cancer.

BRAF V600 mutations are a feasible target for anticancer therapy, occurring in 8% of all solid tumours including substantial numbers of melanomas, thyroid, and 10-15% of colorectal cancers (Davies *et al.* 2002). BRAF mutant tumours use the MAPK/ERK (extracellular signal-regulated kinase signalling cascade) pathway to induce tumour growth (Dienstmann & Tabernero 2011). Specific inhibitors, such as the drug vemurafenib, have shown efficacy against melanoma and slowed cell growth in culture. Specific BRAF inhibition is more effective than non-selective inhibition of the RAF kinase family of the MEK downstream targets (Dienstmann & Tabernero 2011). Vermurafenib is effective for treatment of $BRAF^{V600E}$ mutant melanomas, shown by success in phase I-III clinical trials and FDA approval (Ravnan & Matalka 2012). BRAF inhibitors are effective with regard to efficacy and toxicity; however, acquired drug resistance is a serious problem for this treatment in melanoma, the mechanisms of which need to be explored to ensure optimal clinical outcomes (Dienstmann & Tabernero 2011). Ravnan and Matalka (2012) suggest combination therapy as a solution to vemurafenib resistance in melanoma.

Yang *et al.* (2012) tested vemurafenib as a single-agent $BRAF^{V600E}$ inhibitor in preclinical models of colorectal cancer cell lines and mouse xenografts. Vemurafenib showeddose-dependent inhibition of the MEK and MAPK/ERK signalling pathways which arrested cell proliferation; however, clinical activity was limited in single-agent experiments. Enhanced anti-tumour activity and xenograft survival were found when combined with clinically approved drugs, including EGFR inhibitors, suggesting combination with standard or other targeted treatments is most effective approach for clinical treatment of advanced $BRAF^{V600E}$ mutant colorectal cancers. This is consistent with known association of BRAF mutations with resistance to EGFR inhibitors in colorectal cancer (Di Nicolantonio *et al.* 2008; Yuan *et al.* 2013). Understanding the mechanism of EGFR inhibitor resistance in colorectal cancer is needed for improved application of the drug as a personalised medicine to refine patient selection for single-agent treatment and to develop novel drug combinations; notably, markers of EGFR inhibitor resistance notably include KRAS and BRAF mutations (Loupakis *et al.* 2009; Shaib *et al.* 2013; Siena *et al.* 2009).

Despite successful application of vemurafenib against $BRAF^{V600E}$ in melanomas,

colorectal cancers with $BRAF^{V600E}$ muations have poor prognosis and lack drug response. Prahallad *et al.* (2012) approached this with an RNAi screen of the kinome (a library of 518 genes) in the WiDr colon cell line. Blocking EGFR with shRNA or EGFR inhibiting drugs has shown strong synergy with vemurafenib in cell line experiments; drug synergy was replicated in xenografts. This synergy arises mechanistically from feedback activation of EGFR in response to BRAF inhibition, consistent with higher EGFR expression in colorectal and thyroid cancer cell lines compared to melanoma cell lines and vemurafenib resistance induced from ectopic EGFR expression in melanomas. Corcoran *et al.* (2012) supported this mechanism with transient inhibition of $BRAF^{V600E}$ by vemurafenib which induced rapid reactivation of MAPK/ERK signalling via EGFR in colorectal cell lines. This did not occur in melanoma cell lines suggesting colorectal tissue-specific regulation of EGFR and supporting the use of combined $BRAF^{V600E}$ and EGFR inhibitors which effectively block MAPK/ERK signalling in cell lines and xenografts. Across these studies, synergy between vemurafenib and EGFR inhibitors was replicated for multiple BRAF mutant colorectal cell lines with both antibody (cetuximab) and small-molecule drugs (gefitinib and erlotinib) giving a mechanistically derived means to overcome vemurafenib resistance.

Sun *et al.* (2014) applied these findings to understand vemurafenib resistance in $BRAF^{V600E}$ melanomas. EGFR expression was found to be an adaptive response to BRAF or MEK inhibition in BRAF mutant biopsies and cell lines. An RNAi screen of chromatin regulators (a library of 661 genes) was performed on the A375 melanoma cell line with low initial EGFR expression and selection for vemurafenib resistance. This showed that both gene silencing and drug selection were required for high EGFR expression to emerge. SOX10 was identified as the only gene for which multiple shRNAs could induce EGFR expression and was sufficient for increased EGFR expression and TGFβ signalling by transcriptome analysis. EGFR overexpression or TGFβ treatment also cause a slow growth phenotype, showing oncogene-induced senescence and so are only beneficial to melanomas under drug selection. Therefore, Heterogeneous SOX10 expression levels in melanoma cell lines account for variation in selection and vemurafenib resistance. If BRAF inhibition is removed, SOX10 expression and drug sensitivity are restored giving support to a drug holiday and retreatment approach to counter drug resistance.

Combination treatment of vemurafenib (as a BRAF inhibitor) and various FDA approved EGFR inhibitors entered phase I clinical trials for treatment of $BRAF^{V600E}$ mutant tumours in 2014. Vemurafenib and panitumamab are being trialled in colorec-

tal cancers by the Memorial Sloan Kettering Cancer Center, New York. Combination treatment of vemurafenib, cetuximab (an EGFR inhibitor) and irinotecan (a topoisomerase inhibitor) are being trialled in advanced solid cancers by the MD Anderson Cancer Centre, Houston. Combination vemurafenib and erlotinib are being trialled in BRAF mutant cancers including colorectal and lung by the Peter MacCallum Cancer Centre, Melbourne.

High-throughput Screening for Synthetic Lethality

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

Experimental screening for synthetic lethality is an appealing strategy for wider discovery of functional interactions *in vivo* despite many potential sources of error which must be considered. The synthetic lethal concept has both genetic and pharmacological screening applications to cancer research. Genetic screens, with RNAi to discover the specific genes involved, inform development of targeted therapies with a known mode of action, anticipated mechanisms of resistance, and biomarkers for treatment response. RNAi is a transient knockdown of gene expression more similar to the effect of drugs than complete gene loss and makes comparison to screens in model organisms difficult (Bussey *et al.* 2006). The RNAi gene knockdown process has inherent toxicity to some cells, potential off-target effects, and issues with a high false positive rate. Therefore, it is important to validate any candidates in a secondary screen and

replicate knockdown experiments with a number of independent shRNAs. Alternative gene knockout procedures have also been proposed for synthetic lethal screening including a genome-wide application of the CRIPR/Cas9/sgRNA genome editing technology (Sander & Joung 2014), episomal gene transfer (Vargas *et al.* 2004), or RNAi with lentiviral transfection for delivery of shRNA (Telford *et al.* 2015). Genetic screens have potential for quantitative gene disruption experiments to selectively target overexpressed genes in cancer via synthetic dosage lethality. While powerful for understanding fundamental cellular function, analysis of isogenic cell lines is inherently limited by assuming only a single mutation differs between them despite susceptibility to 'genetic drift' and cannot account for diverse genetic backgrounds or tumour heterogeneity (Fece de la Cruz *et al.* 2015). Genetic screens thus identify targets to develop or repurpose targeted therapies for disease but alone will not directly identify a lead compound to develop for the market or clinical translation.

Chemical screens are immediately applicable to the clinic by directly screening for selective lead compounds with suitable pharmacological properties. However chemical screens lack a known mode of action, may affect many targets, and screen a narrow range of genes with existing drugs. With either approach there are many challenges translating candidates into the clinic such as finding targets relevant to a range of patients, validation of targets, accounting for a range of genetic (and epigenetic) contexts or tumour micro-environment, identifying effective synergistic combinations, enhancers of existing radiation or cytotoxic treatments, avoiding inherent or acquired drug resistance, and developing biomarkers for patients which will respond to synthetic lethal treatment, including integrating these into clinical trials and clinical practice. Identifying specific target genes is an effective way to anticipate such challenges, which can be approached with genetic screens, so we will focus on these and computational alternatives. Screening methods have proven a fruitful area of research, despite being costly, laborious, and having many different sources of error. These limitations suggest a need for complementary computational approaches to synthetic lethal discovery.

Examples of High-throughput Synthetic Lethal Screens Overexpression of genes is another suitable application for synthetic lethality since overexpressed genes cannot be distinguished from the wild-type by direct sequence specific targeted therapy. Overexpression of oncogenes, such as EGFR, MYC, and PIM1, has been found to drive many cancers. PIM1 is a candidate for synthetic lethal drug design in lymphomas and prostate cancers, where it interacts with MYC to drive cancer growth. van

der Meer *et al.* (2014) performed an RNAi screen to for synthetic lethality between PIM1 overexpression and gene knockdown in RWPE prostate cancer cell lines. They recognise RNAi screens are valuable for finding therapeutic targets and biomarkers for therapeutic response. PLK1 gene knockdown and drug inhibition was an effective as a specific inhibitor of PIM1 overexpressing prostate cells in cell culture and mouse tumour xenografts. PLK1 inhibition reduced MYC expression in pre-clinical models, consistent with expression in human tumours which PIM1 and PLK1 are co-expressed and correlated with tumour grade. Therefore PLK1 is justified as a candidate for drug target against prostate cancer progression.

Hereditary leiomyomatosis and renal cell carcinoma (HLRCC) is a cancer syndrome of predisposition to benign tumours in the uterus and risk of malignant cancer of the kidney attributed to inherited mutations in fumarate hydratase (FH). Boettcher *et al.* (2014) performed an RNAi screen on HEK293T renal cells for synthetic lethality with FH. They found enrichment of haem metabolism (consistent with the literature) and adenylate cyclase pathways (consistent with cAMP dysregulation in FH mutant cells). Synthetic lethality between FH mutation and adenylate cyclases was validated with gene knockdown, drug experiments, and replicated across both HEK293T renal cells and VOK262 cells derived from a HLRCC patient, suggesting new potential treatments against the disease. Therefore, synthetic lethality is applicable to metabolic dysregulation in cancer, consistent with the Warburg hypothesis (Warburg 1956), and successfully identifies specific anti-cancer drugs, even when the mechanism is unclear.

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. Aarts *et al.* (2015) screened WiDr colorectal cells for synthetic

lethality between WEE1 inhibitor treatment and an RNAi library of 1206 genes with functions known to be amenable to drug treatment or important in cancer such as kinases, phosphatases, tumour suppressors, and DNA repair (a pathway WEE1 regulates). Screening identified a number of synthetic lethal candidates including genes involved in cell cycle regulation, DNA replication, repair, homologous recombination, and Fanconi anaemia. Synthetic lethality with cell-cycle and DNA repair genes was consistent with the literature and validation in a panel of breast and colorectal cell lines supported checkpoint kinases, Fanconi anaemia, and homologous recombination as synthetic lethal partners of WEE1. These results show that synthetic lethality can be used to improve drug sensitivity as a combination treatment, especially to exploit genomic instability and DNA repair, which are known to be clinically applicable from previous results with BRCA genes and PARP inhibitors (Lord *et al.* 2014). Therefore, WEE1 inhibitors are an example of treatment which could be repurposed with the synthetic lethal strategy and similar findings would be valuable to clinicians as a source of biomarkers and novel treatments. While using a panel of cell lines to replicate findings across genetic background is a promising approach to ensure wide clinical application of validated synthetic lethal partners, a computational approach may be more effective as it could account for wider patient variation than scaling up intensive experiments on a wide array of cell lines and could screen beyond limited candidates from an RNAi library.

Chemical genetic screens are also a viable strategy to identify therapeutically relevant synthetic lethal interactions. Bitler *et al.* (2015) investigated ARID1A mutations, aberrations in chromatin remodelling known to be common in ovarian cancers, for drug response. Ovarian RMG1 cells were screened for drug response specific to ARID1A knockdown cells. They used ARID1A gene knockdown for consistent genetic background, with control experiments and 3D cell culture to ensure relevance to drug activity in the tumour micro-environment. Screening a panel of commercially available drugs targeting epigenetic regulators found EZH2 methyltransferase inhibitors effective and specific against ARID1A mutation with validation in a panel of ovarian cell lines. Synthetic lethality between ARID1A and EZH2 was supported by decreases in H3K27Me3 epigenetic marks and markers of apoptosis in response to EZH2 inhibitors. This was mechanistically supported with differential expression of PIK3IP1 and association of both synthetic lethal genes with the PIK3IP1 promoter identifying the P1I3K-AKT signalling pathway as disrupted when both genes are inhibited. This successfully demonstrates the importance of synthetic lethality in epigenetic

regulators, identifies a therapeutically relevant synthetic lethal interaction, and shows that chemical genetic screens could model drug response and combination therapy in cancer cells. However this approach is limited to finding synthetic lethal interactions between genes with known similar function, which may not be the most suitable for treatment. Further limiting experiments to genes with existing targeted drugs reduces the number of synthetic lethal interactions detected, assumes on their drug specificity to a particular target, and many of these drugs are not clinically available yet anyway as they are still in clinical trials for other diseases or are not supported by healthcare systems in many countries.

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 4x for validated RNAi hits over randomly selected genes. This approach detected known synthetic lethal pairs such as BRCA 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 examples above show that high-throughput screens are an effective approach to discover synthetic lethality in cancer with a wide range of applications. Screens are more comprehensive than hypothesis-driven candidate gene approaches and successfully find known and novel synthetic lethal interactions with potential for rapid clinical application. They have the power to test mode of action of drugs, find unexpected synthetic lethal interactions between pathways, or identify effective treatment strategies without needing a clear mechanism. However, synthetic lethal screens are costly, labour-intensive, error-prone, and biased towards genes with effective RNAi knockdown libraries. Limited genetic background, lethality to wild-type cell during gene knockdown, off-target effects, and difficultly replicating synthetic lethality across different cell

lines, tissues, laboratories, or conditions stems from a high false positive rate and a lack of standardised thresholds to identify synthetic lethality in a high-throughput screen. Therefore there is a need for replication, validation, and alternative approaches to identify synthetic lethal candidates. Varied conditions between experimental screens and differences between RNAi or drug screens makes meta-analysis difficult. Thus genome-scale synthetic lethal experiments are not feasible, even in model organisms, so a computational approach would be more suitable for this task.

Computational Prediction of Synthetic Lethality

Prediction of gene interaction networks is a feasible alternative to high-throughput screening with biological importance and clinical relevance. There are many existing methods to predict gene networks, as reviewed by van Steen (2011) and Boucher and Jenna (2013), summarised in Table 2 below. 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 any of the following 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 the high-throughput two-hybrid screens occur out of cellular context for multicellular organisms.

Table 2. Existing prediction methods for Genetic Interaction Networks in Model Organisms

Method	Input Data	Species	**Source**	**Tool Offered**
Between Pathways Model	PPI, SGI	*S. cerevisiae*	Kelley and Ideker (2005)	
Within Pathways Model	PPI, SGI	*S. cerevisiae*	Kelley and Ideker (2005)	
Decision Tree	PPI, expression, phenotype	*S. cerevisiae*	Wong*et al.* (2004)	2 Hop

Logistic Regression	SGI, PPI, co-expression, phenotype	*C. elegans*	Zhong and Sternberg (2006)	Gene Orienteer
Network Sampling	SGI, PPI, GO	*S. cerevisiae*	Le Meur and Gentleman (2008) LeMeur*et al.* (2014)	SLGI(R)
Random Walk	GO, PPI, expression	*S. cerevisiae* *C. elegans*	Chipman and Singh (2009)	
Shared Function	Co-expression, PPI, text mining, phylogeny	*C. elegans*	Lee*et al.* (2010b)	WormNet
Logistic Regression	Co-expression, PPI, phenotype	*C. elegans*	Lee *et al.* (2010a)	GI Finder
Jaccard Index	GO, SGI, PPI, phenotype	Eukarya	Hoehndorf*et al.* (2013)	
Bimodal Statistics			Wappett (2014)	BiSEp(R)
Machine Learning			Discussed by Babyak (2004) and Lee and Marcotte (2009)	
Machine Learning as discussed by Wu *et al.* (2014)			Qi *et al.* (2008) Paladugu *et al.* (2008) Li *et al.* (2011)	
Machine Learning			Pandey*et al.* (2010)	MNMC
Machine Learning Meta-Analysis			Wu *et al.* (2014)	MetaSL
Flux Variability Analysis Flux Balance Analysis Network Simulation	Metabolism	*E. coli* *Mycoplasma pneumoniae*	Güell*et al.* (2014)	

Table 3. Existing prediction methods for Synthetic Lethality in Cancer

Method	Input Data	Species	**Source**	**Tool Offered**
Network Centrality	PPI	*H. sapiens*	Kranthi*et al.* (2013)	
Differential Expression	Expression, Mutation	*H. sapiens*	Wang and Simon (2013)	
Comparative Genomics Chemical-Genomics	Yeast SGI, Homology	*H. sapiens*	Heiskanen and Aittokallio (2012)	
Comparative Genomics	Yeast SGI, Homology	*H. sapiens*	Deshpande *et al.* (2013)	
Genome Evolution			Lu*et al.* (2013)	
Machine Learning			Discussed by Babyak (2004) and Lee and Marcotte (2009)	
Differential Expression			Tiong*et al.* (2014)	
Literature Database			Li *et al.* (2014)	Syn-Lethality
Meta-Analysis	Meta-Analysis Machine Learning		Wu*et al.* (2014)	MetaSL
Pathway Analysis			Zhang *et al.* (2015)	
Protein Domains	Homology		Kozlov*et al.* (2015)	
Data-Mining			Jerby-Arnon *et al.* (2014) Ryan *et al.* (2014) Crunkhorn (2014) Lokody (2014)	DAISY (method)

Cancer Genome Evolution Hypothesis Test Machine Learning	Expression, DNA CNV, Known SL	*H. sapiens*	Lu*et al.* (2015)	
Chi-Squared Test	Expression, DNA CNV, Methylation, or Mutation	*H. sapiens*	Tom Kelly, Parry Guilford, and Mik Black Dissertation (Kelly 2013) Manuscript in Preparation	SLIPT
Survival Analysis	Expression, Clinical	*H. sapiens*	Mik Black (personal communication)	

There are a number of existing computational methods for predicting synthetic lethal gene pairs. 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. For instance, computational approaches to synthetic lethal prediction are often difficult to interpret, replicate for new genes, or reliant on data types not available for genes that other cancer researchers work on.

Kranthi *et al.* (2013) took a network approach to discovery of synthetic lethal candidate selection applying the concept to 'centrality' to a human PPI network involving interacting partners of known cancer genes. The effect of removing pairs of genes on connectivity of the network was used as a surrogate for viability which is supported by observations that the PPI and synthetic lethal networks are orthogonal in *S. cerevisiae* studies (Tong *et al.* 2004). While they showed the power law distribution expected of a scale-free synthetic lethal network with centrality measures, their approach was limited to known cancer genes and is not applicable to genes without 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 makes using this approach for other cancer genes difficult with respect 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.

A comparative genomics approach by Deshpande *et al.* (2013) used the results of well characterised high-throughput mutation screens in *S. cerevisiae* as candidates for

synthetic lethality in humans (Baryshnikova *et al.* 2010a; Boone *et al.* 2007; Costanzo *et al.* 2010; Costanzo *et al.* 2011; Tong *et al.* 2001; Tong *et al.* 2004). Yeast synthetic lethal partners were compared to human orthologues to find cancer relevant synthetic lethal candidate pairs with direct therapeutic potential. Proposed as a complementary approach to siRNA screens, several synthetic lethal candidates were successfully validated in cell culture; however, this methodology is limited to application on human genes with known yeast orthologues. Synthetic lethal interactions themselves may not be conserved between species (Dixon *et al.* 2009a), although synthetic lethal interactions between pathways may are more comparable. There have been many gene duplications in the separate evolutionary histories of humans and yeast which may lead to differences in genetic redundancy. Yeast are further not an ideal human cancer model because they are do not have tissue specificity, multicellular gene regulation, or orthologues to a number of known cancer genes such as p53.

Differential gene expression has also 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 between tumours with or without functional p53 mutations in Cancer Genome Atlas (TCGA) and Cell Line Encyclopaedia (CCLE) RNA-Seq gene expression data as candidate synthetic lethal partner pathways of p53. Some of these pathways were consistent with the literature and drug sensitivity cell-line screens demonstrating 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. However, the analyses were limited to kinase genes and focused on currently druggable genes, lacking wider application of synthetic lethal prediction methodology. This approach may not be feasible or applicable in cancer genes with a lower mutation rate than p53.

Tiong *et al.* (2014) also investigated gene expression as a predictor of synthetic lethal pairs with colorectal cancer microarrays. Simultaneously differentially expressed "tumour dependent" gene pairs between cancer and normal tissue were used as candidate synthetic lethal interactions. The top 20 genes were tested for differential expression at the protein level with immunohistochemistry staining and correlation with clinical characteristics. Some of the predicted synthetic lethal pairs were consistent with the literature and 2 novel synthetic lethal interactions with p53 were validated in pre-clinical models. While a valuable proof-of-concept for integration of *in silico* approaches to synthetic lethal discovery in cancer, the results again focus on p53 rather

than the wider application of synthetic lethal prediction. The gene expression analyses were conducted in a Han Chinese population with a small sample size (70 tumour, 12 normal) and may not be applicable to other populations.

Another approach to systematic synthetic lethality discovery specific to human cancer (in contrast to the plethora of yeast synthetic lethality data) was to build a database as done by Li *et al.* (2014). In their relational database, called "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 known dedicated to human cancer synthetic lethal interactions and integrates gene functional, annotation, pathway and molecular mechanism data with experimental and predicted synthetic lethal gene pairs. This combination of data sources is intended to tackle the trade-off between more conclusive synthetic lethal experiments in yeast and more clinically relevant synthetic lethal experiments in human cancer models, such as RNAi, especially when high-throughput screens are costly and prone to false positives in either system and difficult to replicate across gene backgrounds. This database centralises a wealth of knowledge scattered in the literature including cancer relevant genes (BRCA, PARP, PTEN, VHL, MYC, EGFR, MSH2, KRAS, and TP53) and is publicly available as a Java App. However, the methodology was not released to replicate or add to the findings with new datasets. Suggested future directions were promising, such as constructing networks of known synthetic lethality, applying known synthetic lethality to cancer treatment, data mining, replicating the approach for synthetic lethality in model organisms, signalling pathways, and develop a complete global network in human cancer or yeast (both of which are still incomplete with experimental data).

From the same group, Wu *et al.* (2014) developed a meta-analysis method (based on the machine learning methods in Table 4) for synthetic lethal gene pairs relevant to developing selective drugs against human cancer. They note that computational approaches scale-up across the genome at lower cost than experimental screens but existing methods are limited by noise and overfitting to a particular predictive feature. Their "metaSL" approach performs well with an AUROC of 0.871 with the claimed strengths of existing machine learning methods and the results are shared on the web. However, once again, the method is not available for analysis of other genes studied by the cancer research community and the method lacks mechanisms, reproducibility, and interpretation by researchers. While machine learning has great potential as

a predictor, it is difficult to interpret which features are being used for prediction and their mechanistic significance, particularly for biologists with limited exposure to computational concepts.

Focusing on the potential for synthetic lethality to be an effective anti-cancer drug target, Zhang *et al.* (2015) used modelling signalling pathways to identify synthetic lethal interactions between known drug targets and cancer genes. A computational approach was again used here to tackle the limitations of experimental RNAi screens such as scale, instability of knockdown, and off-target effects. Strangely, they seemed more concerned with the needs of the pharmaceutical companies than those of the patients. However, their '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 gene expression data (Lee *et al.* 2012) and pathways (the Gene Ontology system). 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. Novel results were enriched for TP53 and AKT synthetic lethal partners, genes known to be important in many cancers but also predicted to be essential by single gene disruption having a large impact on the signalling pathways. Notably, they claim to be able to detect all 3 types of functionally related pathways or protein complexes. The results are consistent with the experimental results in the literature but the group has not shown validation for novel synthetic lethal interactions.

Table 4. Existing Computational Methods used for meta-analysis by Wu *et al.* (2014)

Method	Input Data	Species	**Source**	**Tool Offered**
Random Forest	Machine Learning	*H. sapiens*	Hall*et al.* (2009) Breiman (2001)	WEKA
J48 (decision tree)	Machine Learning	*H. sapiens*	Hall *et al.* (2009)	WEKA
Bayes (Log Regression)	Machine Learning	*H. sapiens*	Hall*et al.* (2009)	WEKA
Bayes (Network)	Machine Learning	*H. sapiens*	Hall *et al.* (2009)	WEKA

PART (Rule-based)	Machine Learning	*H. sapiens*	Hall*et al.* (2009)	WEKA
RBF Network	Machine Learning	*H. sapiens*	Hall *et al.* (2009)	WEKA
Bagging / Bootstrap	Machine Learning	*H. sapiens*	Hall*et al.* (2009)	WEKA
Classification via Regression	Machine Learning	*H. sapiens*	Hall *et al.* (2009)	WEKA
Support Vector Machine (Linear)	Machine Learning	*H. sapiens*	Vapnik (1995)	
Support Vector Machine (RBF – Gaussian)	Machine Learning	*H. sapiens*	Joachims (1999)	
Multi-Network Multi-Class (MNMC)	Machine Learning	*H. sapiens*	Pandey*et al.* (2010)	
MetaSL (Meta-Analysis)	Machine Learning	*H. sapiens*	Wu *et al.* (2014)	MetaSL
Pathway Analysis	Pathway Model	*H. sapiens*	Zhang*et al.* (2015)	

While the mathematical reasoning and algorithms are given, code was not released and it is unlikely that the wider biologically trained research community will be able to reproduce or apply the findings beyond the signalling pathways discussed by Zhang *et al.* (2015). The authors note limitations as directions for further research including the potential of their method to detect mechanisms, types of interactions, impact of activation or inhibition of proteins, and improve performance with a Boolean network or differential equation approach, all of which have been claimed but not shown. Further, this approach is limited by existing pathway data with limited scale, scope, and reliability coming from a range of sources. So far, modelling has been restricted to signalling pathways which are immediately applicable to cancer; while important, the approach lacks broader application to other diseases and pathway types. Zhang et al. (2015) also lack validation, replication, or application of findings and are heavily reliant on existing literature for testing their predictions.

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

predict the candidate synthetic lethal partners of a query gene such as genes recurrently mutated in cancer.

DAISY compares the results of analysis of several data types to predict synthetic lethality, namely: DNA copy number, mutation and gene expression profiles for clinical samples and cell lines. The cell lines data also analysed gene essentiality profiles from shRNA screens. Genes are 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 (Costanzo *et al.* 2010; Kelley & 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. However, this methodology is very stringent, missing potentially valuable synthetic lethal candidates, may not be applicable to genes of interest to other groups and the software for the procedure is not publicly released for replication.

Although the DAISY procedure performs well and has been well received by the scientific community (Crunkhorn 2014; Lokody 2014; Ryan *et al.* 2014), showing a need for such methodology, there is no indication of adoption of the methodology in the community yet. The co-expression analysis may not be the most effective way to test gene expression for directional synthetic lethal interactions (where inverse correlation would be expected). Presumably in the interests of a large sample size, little care is taken to test tissue types separately for tissue specific synthetic lethality (of interest since expression, isoforms, gene function, and clinical characteristics of cancers are 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 reliance of DAISY on co-expression and propose an alternative computational prediction of synthetic lethality based on machine learning methods and a cancer genome evolution hypothesis. Using both DNA copy number and gene expression data from TCGA, a cancer genome evolution model assumes that

synthetic lethal gene pairs behave in 2 distinct ways in response to an inactive synthetic lethal partner gene, either a 'compensation' pattern where the other synthetic lethal partner is overactive or a 'co-loss underrepresentation' pattern where the other synthetic lethal partner is less likely to be lost, since loss of both genes would cause death of the cancer cell. During the cancer genome evolution as the cell becomes addicted to the remaining synthetic lethal partner due to induced gene essentiality. These patterns would explain why DAISY detects only a small number of synthetic lethal pairs, compared to the large number expected based on model organism studies (Boone *et al.* 2007), and the disparity between screening and computationally predicted synthetic lethal candidates due to testing different classes of synthetic lethal gene pairs.

Lu *et al.* (2015) compared a genome-wide computational model of genome evolution and gene expression patterns to the experimental data of Vizeacoumar *et al.* (2013) and Laufer *et al.* (2013). The model had an AUROC of 0.751, performing well for a simpler method than DAISY. 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 14x enrichment of synthetic lethal true positives compared to randomly selected gene pairs. Discovery of such a vast number of cancer-relevant synthetic lethal interactions in humans would not be feasible experimentally and is a valuable resource for research and clinical applications. These predictions are not limited by assuming co-expression of synthetic lethal partners or evolutionary conservation with model organisms enabling wider synthetic lethal discovery. However, there remains a lack of basis for an expectation of how many synthetic lethal partners a particular gene will have, how many pairs there are in the human genome, and whether pathways or correlation structure would influence predicted synthetic lethal partners.

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

To address these needs and concerns raised by recent computational approaches to synthetic lethal discovery in cancer (Jerby-Arnon *et al.* 2014; Lu *et al.* 2015), we propose similar analysis using solely gene expression data which is widely available for a large number of samples in many different cancers. To firmly understand the limitations and implications of synthetic lethal predictions, we propose modelling and simulation of the statistical behaviour of synthetic lethal gene pairs in genomics data. Comparison of synthetic lethal gene candidates from public data analysis, predictions, and networks across datasets will address tissue-specificity concerns. 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.

1.3 E-cadherin as a Synthetic Lethal Target

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

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

Cytoskeleton

Extracellular and Tumour Micro-Environment

Cell-Cell Adhesion and Signalling

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

Stomach Cancers

Breast Cancers

Role in Carcinogensis and Tumourigenesis

Role in Tumour Progression and Metastasis

1.3.3 Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer

Prevalence

Screening, Diagnosis, and Management

Stomach Cancer

Breast Cancer

"Second Hit" Model and Gene Inactivation Mechanisms

1.3.4 Somatic *CDH1* Mutations in Sporadic Cancers

Rate of Mutations

Co-occuring mutations

1.3.5 Models of *CDH1* loss in cell lines

1.4 Summary and Research Direction of Thesis

Synthetic lethality is an important genetic interaction to study fundamental cellular functions and exploit them for biomarkers and cancer treatment. 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, and inconsistent

across independent approaches to different 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. My thesis aims to develop such predictions with a focus on the example of E-Cadherin to compare to the findings of Telford *et al.* (2015) and development of network approaches to tissue specificity with the following bioinformatics and computational biology investigations:

- Simulate gene expression data, construct a statistical model for synthetic lethality, and measure performance of testing for synthetic lethal genes.
- Apply Synthetic lethal prediction to public genomics data
- Gene expression
- DNA copy number
- DNA methylation
- Somatic Mutation
- Select candidates for synthetic lethality with CDH1 in breast cancer, compare to RNAi data for validation, and triage candidates for drug development against HDGC and sporadic breast cancers
- Release a synthetic lethal prediction methodology to the research community for wider application
- Construct and analyse genome-scale synthetic lethal networks, tissue specificity, or drug response using synthetic lethal predictions

Chapter 2

Methods, Techniques, and Resources

2.1 Overview/meta-text

2.2 Bioinformatics to Enable Genomics Research

2.2.1 Public Data and Software Packages

Bioinformatics resources, such as databases and methods, have become an integral part of genetics and genomics research. Reference genomes are used routinely to facilitate more effective experiments and for mapping reads from later genomics and transcriptomics studies. These include studies of large-scale genetic variation, including risk factors for disease, and gene expression studies, including those in cancers.

Similarly, various gene expression databases have been developed for sharing gene expression data from previous studies, including Gene Expression Omnibus (GEO), caArray, and arrayExpress. These were originally developed to share microarray gene expression data but now many support RNA-Seq data (with the benefits discussed previously) and have set a precedent for data sharing, data mining, and the wider benefits of publicly available data for enabling the scientific community further utilise the data compared to a single research group or consortium. Such practices of integrating findings from publicly available genomics data with the research questions and experimental results of individual research groups has carried over into RNA-Seq datasets including the large-scale cancer genomics projects (such as TCGA). The thesis is one such example of an investigation enabled by this wider movement and tools developed

in various disciplines to generate, disseminate, and process genomic-scale data.

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

[**more detail e.g., TCGA, Reactome**]

2.2.2 Computational Tools and Enabling Biological Research (remove/state assumptions only)

There remains debate regarding the optimal methods to perform an alignment. However, the statistical and biological aspects of Bioinformatics are the focus of this thesis, comparing alignment methods is outside the scope of this investigations. The TCGA project used the widely adopted "Bowtie" tools for alignment, with "mapslice" to detect splice sites, and the Reads Per Kilobase per Million mapped reads (RPKM) approach to qualify reads per transcript as a measure of gene expression. These are widely acceptable tools for processing RNA-Seq data and this is the raw data publicly available from TCGA.

High Performance and Parallel Computing Another significant development in computer science for bioinformatics is parallel computing, performing independent operations in separate cores, such "multithreading" is widely used to increase the time to compute results. Bioinformatics is particularly amenable to this since performing multiple iterations of a simulation or testing separate genes is often "embarrassingly parallel", being completely independent of the results of each other. As such parallel computing is offered by many high-performance "supercomputers" including national research infrastructure. The New Zealand eScience Infrastructure (NeSI) is once such computational resource providing the Intel Pan cluster (hosted by the University of

Auckland) used throughout this thesis project to optimise and perform computations which would have otherwise been infeasible in the timeframe of thesis. This is another example of how technological developments and infrastructure has enabled research including this project.

2.2.3 Gene Expression Analysis and Statistical Challenges

Hypothesis Testing and Multiple Comparisons Procedures

Candidate Triage and Integration with Experimental Data

2.2.4 Mathematical Challenges in Bioinformatics

Graph Theory, Systems, and Network Biology

Matrix Operations and Pathway Metagenes

2.3 Data Handling

2.3.1 Normalisation (voom)

2.3.2 Sample triage

2.3.3 SVD/mg

2.4 Techniques

2.4.1 Clustering

2.4.2 Heatmap

2.4.3 Modeling and Simulations

(AU)ROC

2.4.4 Permutation / Resampling

2.4.5 Network Metrics / Techniques

Network theory is an interdisciplinary field which combines the approaches of in computer science with the metrics and fundamental principles of graph theory, an area of

pure mathematics dealing with relationships between sets of discrete elements. The first large networks were generated randomly and exhibited interesting small-world and scale-free properties. Application of network theory in the life sciences has, until recently, been largely restricted to small networks in sociology or ecology. The vast amounts of molecular and cellular data from high-throughput technologies have raised the need for systems-level, network-based, and genome-wide bioinformatics analysis to capture the complexity of a cell at the molecular level and understand aberrations in cancer.

Graph theory is a branch of pure mathematics which deals with the properties of sets of discrete objects (referred to as a 'node' or 'vertex') with some pairs are joined (by a 'link' or an 'edge'). Originally conceived as a reductionist abstraction to solve problems in mathematics and more complex problems later in computer science, graph theory serves as the fundamental basis for a wide range of studies including material physics, traffic analysis, computer architecture, and phylogenetic trees. Applications vary depending on the situation modelled, particularly in how the edges between vertices are defined, whether they are directed or weighted, and whether multiple redundant edges between a pair of vertices (referred to as 'parallel edges') or edges connecting a vertex to itself (referred to as 'loops') are permitted in the model. Networks are defined such that the edges represent a relationship between the vertices and may be directed, weighted, or contain parallel edges or loops depending on the application.

Network theory is the sub-discipline of graph theory which deals with networks which has become popular due to the vast potential for applications of networks. The properties of large networks were studied by constructing random networks by randomly linking a fixed number of nodes (Erdős & Rényi 1959; Erdős & Rényi 1960). Despite the random nature of these networks, properties such as their connectivity were well characterised. The vertex degree 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 & Oltvai 2004).

This work formed the foundation for studying complex networks 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 (Milgram 1967; Travers & Milgram 1969), is the remarkably short path lengths between any nodes in a small world network. A small world network is well-connected with a characteristic path length proportional to the logarithm of thenumber of nodes (Watts & Strogatz 1998). Watts and Strogatz (1998) developed a model of random rewiring of a regu-

lar network to construct random networks with the small world property and a high clustering coefficient. While these properties are more representative of networks occurring in nature, their model is limited by the degree distribution which converges to a Poisson distribution as it is rewired (Barrat & Weigt 2000).

The degree distribution of naturally occurring networks often follows a power law distribution with the majority of nodes having far fewer connections than average and a small subset of highly connected network 'hubs'. Barabási and Albert (1999) 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. The Barabási and Albert (1999) scale-free networks are limited by a low clustering coefficient and lack of modular structure; however, they have enabled the study of scale-free network topology and served as a basis for modified scale-free models (Dorogovtsev & Mendes 2003; Holme & Kim 2002).

Han *et al.* (2004) observed dynamic modularity in biological networks and suggested the network structure may underpin genetic robustness and plasticity. They focus on network hubs which are more likely to be essential genes and define the subgroups of hubs based on correlation of gene expression with protein-protein interaction partners: 'party' hubs (which interact simultaneously with many partners) and 'date' hubs (which interact with different partners in different conditions). Party and date hubs occurred most frequently within and between network modules respectively. Party hubs were considered local regulators, whereas date hubs were considered important to network connectivity as global regulators. This distinction between classes of network hubs was supported by differences in tissue specificity and clinical relevance as a proposed predictor of clinical outcome in breast cancer with an AUROC of 0.784 (Taylor *et al.* 2009). However, correlation between expression and protein interactions were not robustly reproduced. The importance of date hubs has been criticised for assuming a bimodal distribution and basing the global importance of data hubs on a small subset (Agarwal *et al.* 2010). As an alternative interpretation, Agarwal *et al.* (2010) suggest the importance of interactions rather than network hubs as interactions important to

the network were between functionally similar proteins. Network hubs can also be classed as associative or dissociative depending on whether they tend toward or away from connecting directly to other network hubs (van Steen 2010). The associative and dissociative properties can also be used to test whether nodes of a particular subgroup (e.g., gene function) associate with each other.

Applications of network theory are diverse, including uses in social sciences, engineering, and computer science. Due to their complexity and difficulty of gathering sufficient empirical data, biological applications of network theory are relatively unexplored. High-throughput technologies such as siRNA screens, two-hybrid screens, microarrays and massively parallel sequencing have made generating genome-scale molecular data feasible and enabled analysis of biological networks at the molecular level. Many types of inter-related molecular networks can be constructed and analysed, depending on the biological application, the way interactions between molecular components are defined and the data used to generate them as shown in Table 1.

Table 1. Types of biological network based on molecular data

Network Type	Node	Interaction	Data Generation
Co-expression	Gene	Correlation of expression	Array, RNA-Seq
Protein (Physical)	Protein	Protein-protein binding	Two-Hybrid
Signal Transduction	Protein	Protein mediated signalling	Curate known pathways
Metabolic	Metabolite or cofactor	Involved in same reaction (links by reactions/enzymes)	Curate known pathways
Chemical or Drug	Protein	Targets of the same drug	Curate known drug targets
Regulation	Gene	Regulate each other by encoded proteins	Array, RNA-Seq, ChIP
RNA or DNA binding	Gene or miRNA	Binding of encoded RNA or protein or DNA or mRNA	ChIP, RIP
Functional	Gene	Shared gene function	Curate known pathways
Genetic Interaction	Gene	Unexpected phenotype with combined loss of function	SGA, EMAP, DAmP, siRNA

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.

2.5 Pathway Structure Methods

2.5.1 Sourcing graph structure data

2.5.2 Constructing pathway subgraphs

2.5.3 Centrality Measures

2.5.4 Data Sources

As summarised in Table 5, there is are a vast array of publicly available resources for model organism, human, and cancer genomics analysis. These will be used as a resource for bioinformatics analysis of genetic interactions and networks in addition to modelling and simulation approaches.

Table 5. Data sources for Research

Database		**Study Type**	**Data Types Supported**
TCGA	Cancer Genome Atlas	Cancer	Sequence, Mutation, CNV, SNP, expression, DNA Methylation, Clinical, RNA-Seq
ICGC	International Cancer Genome Consortium	Cancer	Sequence, Mutation, CNV, SNP, expression, DNA Methylation, Clinical, RNA-Seq
ENCODE	Encyclopaedia of DNA Elements	Human Normal Tissue Cancer Cell Line	Sequence, expression, DNA/RNA binding RNA-Seq, ChIP-Seq, RIP-Seq
CCLE	Cell Line Encyclopaedia	Cancer Cell Line	DNA CNV, expression, mutation, drug sensitivity
GEO	Gene Expression Omnibus	Various	Gene Expression, RNA-Seq

GENT	Gene Expression Atlas	Human Normal Tissue and Cancer	Gene Expression
BIND	Biomolecular interaction network database	Model Organisms Human	DNA, RNA, ligand, or protein binding Two-hybrid data, ChIP, CLIP, RIP
STRING	Search Tool for retrieving interacting genes/proteins	Various	Protein-protein interactions: curated and predicted from MINT, PPRD, BIND, DIP, BioGRID, KEGG, Reactome, IntAct, EcoCyc, NCI, and GO
BioGRID	Biological General Repository for Interaction Datasets	*S. cerevisiae*, *S. pombe*, *A. thaliana*, etc	Protein and Genetic Interaction
Human Interactome	Human Interactome	Human	Protein Interactions
DRYGIN	Data Repository of Yeast Genetic Interactions	Yeast	Genetic Interactions
SGD	Saccharomyces Genome Database	Yeast	Sequence, Expression, Protein and Genetic Interactions
GO	Gene Ontology	Various	Gene Function
KEGG	Kyoto Encyclopedia of Genes and Genomes	Various	Gene Function and Pathway
Reactome		Various	Gene Function and Pathway
DrugBank		Human	Chemical and drug target sequence/structure
TTD	Therapeutic Target	Human	Protein and Nucleic Acid drug targets
TDR	Targets database	Human	Chemical Genomics (Tropical Disease)
BindingDB			Protein-drug binding affinity

Chapter 3

Methods Developed During Thesis

3.1 Overview/meta-text

3.2 Developing a Synthetic Lethal Detection Methodology

3.2.1 Rationale and Design of Test

3.2.2 Synthetic Lethal Detection Method

3.3 Simulations and Modelling Synthetic Lethality in Expression Data

3.3.1 Synthetic Lethal Modeling

3.3.2 Simulation Procedure

3.4 Assessing the Synthetic Lethal detection methodology (simulation results part 1)

3.4.1 Binomial Simulation of Synthetic lethality[?]

3.4.2 Multivariate Normal Simulation of Synthetic lethality[+dir?]

Receiver Operating Characteristic Curves

Simulated Expression Heatmaps

3.4.3 Replication Simulation Heatmap

3.5 Graph Structure Methods

Chapter 4

Synthetic Lethal Analysis of Gene Expression Data

4.1 Abstract

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

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

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

used to predict known genetic interactions.

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

4.2 Aims and Significance

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

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

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

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

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

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

4.3 Background

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 Background

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

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

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

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

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

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

4.5 Sourcing TCGA data

4.6 Quality checking

4.7 Global Synthetic Lethality

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

4.8 CDH1 Analysis with Subgroups

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

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

4.9 Cell Line Analysis

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

4.10 Mutation, Copy Number, and Methylation

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

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

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

4.11 ANOVA of Expression Predictors

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

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

led to investigations into the simulation of synthetic lethality.

4.12 Mutation Analysis, Pathway Expression, and Metagene Synthetic Lethality

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

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

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

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

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

4.13 Data clean up, gene SL, and pathway SL

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

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

(Olszanski 2014).

4.14 Overview of Challenges

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

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

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

HDGC patients.

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

4.15 Comparison of gene SL predictions and siRNA screen candidates

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

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

screen data.

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

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

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

4.17 Comparison of candidate SL Pathways

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

4.18 Future Directions

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

4.19 Hub Genes

4.20 Metagene pathway expression

4.21 Metagene synthetic lethality

4.22 Replication in stomach cancer

4.23 Important Results

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

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

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

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

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

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

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

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

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

The expression profiles of the SL partners of CDH1 predicted from the TCGA breast cancer RNA-Seq data in CDH1 low tumours (where synthetic lethal partners are expected to have compensating high or stable expression) are shown in Figure 7

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

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

Chapter 5

Pathway Structure of Synthetic Lethal Genes

5.1 Abstract

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

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

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

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

5.2 Background

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

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

5.4 Synthetic lethal genes in synthetic lethal pathways

5.5 Methods

5.5.1 Sourcing graph structure data

5.5.2 Constructing pathway subgraphs

5.5.3 Centrality Measures

5.5.4 upstream and downstream gene detection

5.5.5 permutation analysis

5.6 Centrality and connectivity of synthetic lethal genes

5.7 Upstream or downstream synthetic lethal candidates

5.8 Hierachical approach

5.9 Discussion

5.10 Conclusion

Chapter 6

Simulation and Modeling of Synthetic Lethal Pathways

6.1 Abstract

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

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

6.2 Background

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

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

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

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

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

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

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

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

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

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

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

6.3 Simulations and Modelling Synthetic Lethality in Expression Data

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

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

6.4 Developing a Synthetic Lethal detection methodology

6.4.1 Testing Multivariate Normal Simulation of Synthetic lethality

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

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

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

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

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

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

6.4.2 Receiver Operating Characteristic Curves

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

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

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

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

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

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

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

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

6.4.3 Simulated Expression Heatmaps

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

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

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

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

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

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

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

6.4.4 Replication Simulation Heatmap

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

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

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

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

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

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

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

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

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

6.5 Simulation of synthetic lethality in graph structures

6.5.1 Developing a multivariate normal expression from graph structures

6.5.2 Simulations over simple graph structures

Performance

Synthetic lethality across graph stuctures

Performance with inhibition links

Performance with 20,000 genes

6.5.3 Simulations over pathway-based graphs

6.5.4 Comparing methods

SLIPT and Chi-Squared

Correlated query genes

Correlation

Linear models

6.5.5 Developing a linear model predictor of synthetic lethality

Linear models

Polynomial models

Conditioning

SLIPTv2

6.6 Significance

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

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

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

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

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

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

6.7 Future Directions

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

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

research community (manuscript and code release in preparation).

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

6.8 Conclusion

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

Chapter 7

Discussion

This is the first chapter. Here is an example citation: Rountree (1998).

Chapter 8

Conclusion

References

Rountree, N. (1998). *How to do Anything.* Dunedin, New Zealand: Non-existent Publishers.